Generalized Prediction Model for Detection of Psychiatric Disorders

by

Bhaskar Sen

A thesis submitted in partial fulfillment of the requirements for the degree of

Master of Science

Department of Computing Science

University of Alberta

Abstract

Computer aided diagnosis of mental disorders like Attention Deficit Hyperactivity Disorder (ADHD) and Autism is a primary step towards automated detection and prognosis of these psychiatric diseases. This dissertation applies analyses based on learning models that use structural texture and functional connectivity to diagnose ADHD and also Autism, from (structural) 3-dimensional magnetic resonance imaging (MRI) and 4-dimensional resting-state functional magnetic resonance imaging (fMRI) scans of subjects. One model learns texture-based filters that are used to extract features from MRI scans. Using these learned features, the model achieves 0.6257 (baseline 0.5497) accuracy on the ADHD-200 hold-out dataset for differentiating between healthy control vs ADHD patients and also achieves 0.6173 (baseline 0.5157) accuracy on the ABIDE (Autism) hold-out test for differentiating between healthy control vs Autism patients. Our next model examines temporal sequence of fMRI activation levels at various brain locations in order to make a diagnosis from fMRI scans. This incorporates spatial nonstationary independent component analysis of the fMRI scans in order to extract the uncorrelated components and decomposes fMRI scans into common spatial components and corresponding time courses. Using individual time courses of 45 independent components as features, our algorithm learns a classifier that yields an accuracy of 0.6491 on the ADHD-200 hold-out dataset, and 0.6233 accuracy on the ABIDE hold-out test. This result is higher (0.0231 for ADHD and 0.0233 for Autism) than previously published accuracies on these datasets using fMRI scans. Finally a combination of multimodal features yields 0.6725 diagnosis accuracy on ADHD-200 and 0.6431 accuracy on ABIDE. This result is significantly higher (0.0465 for ADHD with one sided p = 0.01 and 0.0431 for Autism with one sided p = 1.6172e-06) than previously published hold-out accuracies on these datasets using only imaging data. Our results indicate that combining multimodal features yields good classification accuracy for diagnosis of ADHD and Autism, which is an important step towards computer aided diagnosis of these psychiatric diseases.

To my parents

For teaching me everything I know

What we observe is not nature itself but nature exposed to our method of questioning. Our scientific work in physics consists in asking questions about nature in the language that we possess and trying to get an answer from experiment by the means that are at our disposal.

– Werner Heisenberg, 1958.

Acknowledgements

Foremost, I would like to express my sincere gratitude to my advisors Prof. Russell Greiner and Dr. Matthew Brown for their continuous support of my MSc. study and research, for their patience, motivation, enthusiasm, insightful comments, hard questions and immense knowledge. Their guidance helped me in all the time of research and writing of this thesis. I thank my fellow labmates at the University of Alberta, Zheng Shi, Roberto Vega, Ping Jin, Luke Kumar, Graham Little and Mina Gheiratmand for helpful discussions during the course of study. Last but not the least, I would like to thank my family: my parents and grandmother for supporting me spiritually throughout my life.

Table of Contents

List of Tables

List of Figures

Chapter 1

Introduction

Statistical machine learning methods have recently permeated disciplines such as Psychiatry, which specializes in the diagnosis and treatment of neuropsychiatric disorders [9]. The availability of large scale neuroimaging datasets has encouraged researchers to develop computer aided tools and procedures for understanding the human brain and its disorders. Structural MRI scans provide a noninvasive technique for getting a volumetric image of the brain anatomy. On the other hand, functional MRI (fMRI) scans measure brain activity by detecting fluctuations in blood-oxygen levels over time.

Using MRI/fMRI for detecting brain functional disorders, like Autism and Attention Deficit Hyperactivity Disorder, remains an unsolved challenge for neuroscientists. Significant work has been done to show changes in brain connectivity for the patients suffering from ADHD [21, 22, 23] and Autism [19, 20]. These association studies, which find specific characteristics that are discriminating at the group level (patient vs control), are very useful as a primary step towards robust understanding of the diseases and their underlying differentiating factors. By contrast, we are interested in exploring ways to learn *predictive models*, which seek combination of features that are effective for predicting whether an individual subject has the disease.

Here we explore ways to use analyses based on structural texture features (from structural MRI scans) and functional connectivity (from fMRI scans) to predict whether a specified subject has specific psychiatric disease. In both cases, brain structure (MRI) and activities (fMRI) are represented digitally as *voxels*, the smallest block in the MRI and fMRI scans. Each voxel in MRI and fMRI corresponds a 3-dimensional rectangular block in brain. The size of that block depends on the resolution of the scanner. A high resolution scanner will be able to take images with very small sized voxels.

In brain images, the structural textures give us information about spatial arrangements of voxel intensities in the 3D scans which in turn describes neurological aspects of the subject's brain. On the other hand, functional connectivity captures patterns of deviations from statistical independence between the time signals at distributed and often spatially remote neuronal regions (Friston, 1993; 1994, Sporns 2010). Deviations from statistical independence are generally taken to indicate dynamic coupling and can be measured, for example, by estimating the correlation, independent components, etc. As a first step towards creating a generalized prediction tool for ADHD/Autism, we experiment with the texture based models that describe structural arrangements of the brain by learning texture features from 3D MRI scans. We then use these texture features to learn classifiers designed to predict whether a patient has the disease. To study the functional aspects of the diseases, we experiment with four different source separation techniques to decompose the fMRI into spatial components (each of the components consists of a set of voxels in brain that co-activate together) that are common across the subjects, with individual time courses. This in turn allows each subject to be described as a composition of the common spatial maps with unique time course corresponding to that subject. Instead of doing statistical group level analysis on the extracted time courses, we use them as features for our predictor. We used two publicly available multisite datasets, ADHD-200 [1] and ABIDE [2], for testing our models.

Since the publication of ADHD-200 data, many researchers have explored ways to improve the prediction accuracy of ADHD using this data. However, the competition results show that the best prediction result is still far from being clinically relevant. Eloyan et al. [3] won the ADHD-200 competition (using imaging data), achieving an accuracy of 0.6154 (baseline 0.5497). Later other works [13, 14, 10] improved the accuracy on this task, with an accuracy 0.6257 [10] on the hold-out set. Similarly, using ABIDE data for autism prediction, Nielson et al. [18] achieved 0.6000 (baseline 0.5157) prediction accuracy.

In this dissertation, we show that by using multimodal features from both structural and functional MRI scans, we can improve the accuracy of ADHD/Autism prediction using MRI/fMRI scans compared to the previous results. This is an important step towards formulating a computer aided prediction model for predicting psychiatric diseases.

Specific contributions of this work are as follows:

1. Motivated by the work of Ghiassian et al. [10] of using texture-based features for ADHD/Autism prediction, we show that texture based models that learn features from the data, can match, or sometimes outperform, other standard MRI-based prediction models (Section 2.4).

2. We extend existing ICA-based source separation in fMRI analysis to prediction studies and devise a novel algorithm for this task (Section 2.5). This algorithm outperforms other fMRI based algorithms for ADHD/Autism prediction.

[1] `http://fcon_1000.projects.nitrc.org/indi/adhd200/`

[2] `http://fcon_1000.projects.nitrc.org/indi/abide/`

3. We apply and compare four different source separation models to get a good representation of fMRI scans and validate each model based on the prediction accuracy on a large ADHD (respectively Autism) dataset.

4. We introduce using multiple decorrelation [7] for fMRI source separation. Here, 3D volumes of fMRI are modeled as nonstationary signals and source separation based on second order criteria is used to separate out common spatial activation maps and corresponding time courses. The validation set and test set accuracies show that there are changes in the components' time courses that distinguish ADHD (respectively Autism) patients' brains from healthy controls.

5. Finally, combining features learned from MRI and features extracted using fMRI source separation, we improve the prediction model accuracy for ADHD to 0.6725 (compared to 0.626) and Autism to 0.6431 (compared to 0.60). Note that these results are just using imaging data.

In order to develop a classifier that can diagnose ADHD and Autism correctly, we took a *"biologically naive"* approach for extracting features; that is, we do not use any prior biological information, about the brain nor the fMRI signal, etc. The ADHD-200 and ABIDE datasets consist of both structural and functional MRI scans. Figure 1.1 shows the pipeline for developing the classification algorithm. We develop separate diagnostic classifiers for diagnosis from MRI (**Method 1** and **Method 2**) and fMRI scans (**Method 3**). Finally, we used a combination of features from both MRI and fMRI scans to develop the *"final"* classifier (**Method 4**). Fig. 1.2 further elaborates the feature extraction block in Fig. 1.1 for each of these methods. In all these cases, we assume that the features are separable in a *non-linear* space. When we use *linear* features − i.e. features are extracted using a linear transform − we use a *non-linear* classifier. In case of *non-linear* features − i.e. features are extracted using a non-linear transform − we use *linear* classifier.

For any image recognition task, the choice of feature extraction can greatly facilitate or impede the classification. If the features are meaningful and discriminating, the task for the subsequent classifier becomes easier [37, 39, 41, 42]. For developing a diagnostic classifier from structural MRI scans, we consider two approaches.

First, motivated by recent success of generic feature descriptors extracted from convolutional neural networks in object recognition and classification tasks [37, 38], we apply the large generic filters learned from a vast labeled dataset (ImageNet[3]). We used the `Overfeat` system [4] [38], which has a vast array of filter banks learned from ImageNet, to our learning task (predicting ADHD vs. Healthy and Autism vs. Healthy). Here, features were extracted using the filter banks in `Overfeat` system to apply in totally different domain and

[3] `http://www.image-net.org/`

[4] `http://cilvr.nyu.edu/doku.php?id=software:overfeat:start`

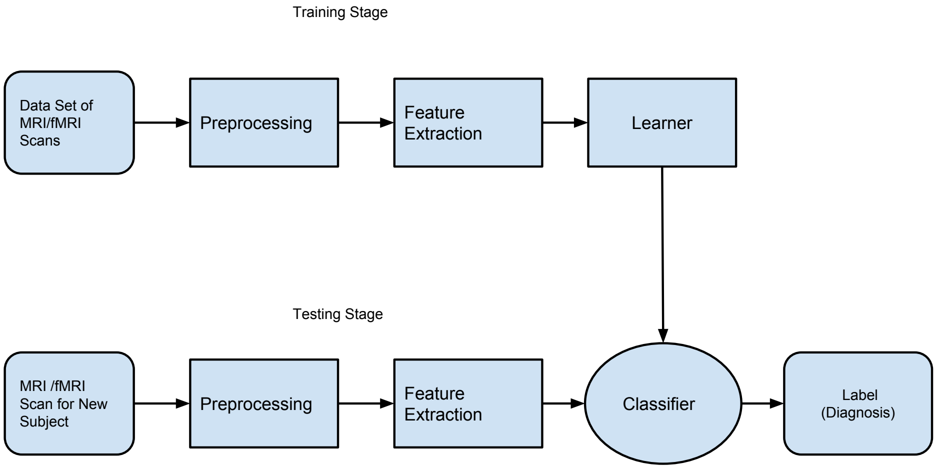


Figure 1.1: Overview of our diagnosis models. Each model's training stage develops the classifier responsible for diagnosing new subjects, which is then used in testing stage

datasets (ADHD-200 and ABIDE) [39]. We denote this model for learning and classification as **"Method 1"**. Note that **"Method 1"**(and the other methods discussed below) include both learning and classification.

Secondly, many current projects in machine learning and neuroimaging research have attempted to learn features from the data [31], which boost the prediction power of the model. There is a vast literature that employs domain independent representation learning (that learns features from the data) for many recognition and prediction tasks in computer vision. These feature learning techniques have recently been employed for learning features for Alzheimer's disease from the ADNI dataset [32]. In these datasets, the number of labeled MRI/fMRI scans from patients and controls is on the order of hundreds or thousands for the largest public datasets. This is relatively small in comparison to datasets such as ImageNet, which has over half a million samples. MRI/fMRI datasets are comparatively small due to the difficulties of recruiting subjects and the high costs associated with MRI scanning. Therefore, it is imperative to devise tools that explore the local structures and redundancies in the images. Motivated by the success of using representation learning in medical imaging domains [32], we used a one layer convolutional neural network with 3-D learned kernels ($5 \times 5 \times 5$) as feature extractors using structural magnetic resonance images from each

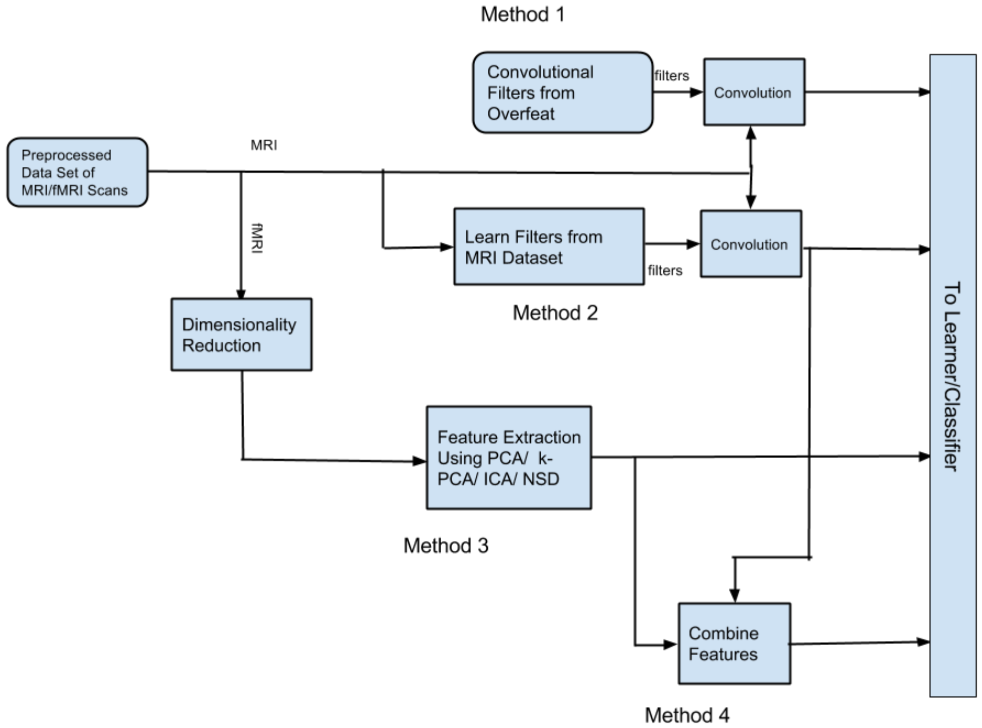

Method 1
Convolutional Filters from Overfeat
filters
Convolution
Preprocessed Data Set of MRI/fMRI Scans
MRI
fMRI
Learn Filters from MRI Dataset
filters
Convolution
Method 2
Dimensionality Reduction
Feature Extraction Using PCA/ k-PCA/ ICA/ NSD
Method 3
Combine Features
Method 4
To Learner/Classifier

Figure 1.2: Feature extraction steps for each of our methods

dataset (ADHD-200/ ABIDE). The main idea of using self-taught learning is motivated by the work of Raina et al. [45]. It uses sparse coding to construct higher-level features using the unlabeled data. These features form a succinct input representation and significantly improve classification performance for object recognition. In this experiment, our main contribution is to show that a simple feature learning algorithm can sometimes perform as well as a complex algorithm. We denote this method for learning convolutional network features and classification from MRI scans as **"Method 2"**.

For developing a diagnostic classifier from fMRI scans, we consider four different source separation techniques for learning a good representation of fMRI scans. Each of the models uses the temporal evolution of fMRI voxel activations in the brain in order to make a diagnosis. It decomposes fMRI scans into common spatial components and corresponding time courses. Specifically we apply *Principal Component Analysis (PCA)*, *Kernel Principal Component Analysis (k-PCA)*, *Independent Component Analysis (ICA)* and *Nonstationary Source Decomposition (NSD)* to the problem of learning to diagnose ADHD (respectively Autism) from resting-state fMRI scans of subjects. The individual time courses from the separated sources were used as features for learning and classification. We denote this method for learning and classification using four different feature extraction approaches from fMRI scans as **"Method 3"**. Note that **"Method 3"** refers to each of four separate methods, each of which is one of PCA, kPCA, ICA and NSD.

Finally, we combined different features from imaging modalities (MRI and fMRI) to create the final predictor. Recent neuroimaging studies have indicated higher predicting capability from combined features from different neuroimaging modalities [47, 49]. To investigate the effect of combining features from structural MRI and functional MRI on psychiatric disease prediction, we concatenated the features from previous experiments (**"Method 2"** and **"Method 3"**). The combined features were used for the learning and prediction. We denote this method as **"Method 4"**.

The rest of the thesis is structured as follows: Chapter 2 outlines the pre-processing of raw fMRI data and overviews the methods used in our study. Chapters 3 describes the results on the ADHD-200 and ABIDE datasets, and Chapter 4 discusses potential future works for MRI and fMRI-based diagnosis.

Chapter 2

Foundation

This section presents the overall process of the diagnostic system based on Figure 1.1. We first describe the dataset (Section 2.1) and evaluation criteria (Section 2.2). Then we outline the preprocessing pipeline for ADHD-200 and ABIDE data in Section 2.3. The remaining sections summarize diagnostic methodology from MRI scans (Section 2.4), fMRI scans (Section 2.5) and combined imaging features (Section 2.6).

2.1 Dataset

For evaluating each model, we used two multi-site datasets: ADHD-200 and ABIDE. Each of the datasets included a structural scan (high resolution, for a single time point), and also one or more resting-state functional scans for each of the subjects. Spatial resolution of the structural MRI scans was $1mm \times 1mm \times 1mm$. In the resting state functional scan, the subject did not perform any explicit task. That functional scan included between 76 to 261 time points for each ADHD-200 subject and between 82 and 320 time points for each ABIDE subject. Different subjects were scanned with different temporal resolutions: ranging from 1.5 seconds through 3 seconds in the ADHD-200 dataset, and from 1 seconds through 3 seconds in the ABIDE data. The field strength of the MRI scanners varied from 1.5T to 3T. Each data collection site used its own scanner(s) and its own MR scanning parameters. More details are available at the ADHD-200 site [1] and ABIDE site [2] .

ADHD-200

The ADHD-200 data is a multi-site combination of neuroimages taken from 8 sites. The demographics of subjects in the dataset is shown in Table 2.1.

We used a training set from this data to learn a prediction model and to estimate its accuracy (using cross-validation). The training set consists of 776 resting state scans: 491 were taken from healthy controls and 279 were patients. To balance our training set, we

[1] `http://fcon_1000.projects.nitrc.org/indi/adhd200/`

[2] `http://fcon_1000.projects.nitrc.org/indi/abide/`

Table 2.1: ADHD-200 data demographics. Site abbreviations: *Peking University (Peking), Kennedy Krieger Institute (KKI), NeuroIMAGE (NI), New York University (NYU), Oregon Health and Science University (Oregon), University of Pittsburgh (Pitt), Washington University in St. Louis (WashU)*

	Peking	Brown	KKI	NI	NYU	Oregon	Pitt	WashU
Subjects	245	26	94	73	263	113	98	61
ADHD	130	26	33	50	163	71	9	0
Male/Female	174/71	9/17	64/30	43/30	171/92	61/52	53/45	33/28
Age Mean	11.7	14.54	10.22	17.64	11.45	0.10	15.08	11.47
Age STD	1.96	2.54	1.34	3.05	2.91	1.20	2.78	3.88

used all 279 patients and selected 279 healthy controls evenly taken from all the sites as our training set [14, 40, 13] which means that the baseline classification accuracy for the training set is 0.50. This training set is used for model selection and cross validation. The ADHD-200 competition hold-out data consists of 171 subjects (94 healthy subjects and 77 ADHD cases, baseline 0.5497). The set is used for evaluating the quality of the final model which was untouched during training. We also discuss the effect of unbalanced set on training (in terms of number of patients vs healthy) in section 3.4.1. The ADHD-200 dataset included other non-imaging features for each subject, including gender, age, handedness, site of the imaging, IQ measure etc (see ADHD-200 consortium[3] and Brown et al. [50] for more details on these personal characteristics). However we only use imaging data for our experiments.

ABIDE

The ABIDE[4] dataset consists of 1111 scans: 573 are healthy controls and 538 patients with autism. The demographics of subjects in the dataset is shown in Table 2.2.

To evaluate each learning model, we used 800 subjects (70%) for model training and 311 subjects (30%) for hold-out testing. We used the same case/control ratio (0.5157) for both training and test set. The ABIDE dataset provided an extensive array of nonimaging information information which included age, gender, handedness, various IQ scores, site of the imaging and eyestat (which indicated whether the person kept his eyes open or not during the scan); for more information on these personal characteristics see ABIDE [4]. Again we only use imaging data for our experiments.

2.2 Evaluation Criteria

We use both 5-fold cross validation accuracy and hold-out accuracy to evaluate our implemented diagnosis algorithms. Five-fold cross validation is mainly used for tuning each model and getting a basic estimate of performance for the model. The training set S and

[3] `http://fcon_1000.projects.nitrc.org/indi/adhd200/`

[4] http://fcon_1000.projects.nitrc.org/indi/abide/

Table 2.2: ABIDE data demographics. Site abbreviations: *California Institute of Technology (Caltech), Carnegie Mellon University (CMU), Kennedy Krieger Institute (KKI), Ludwig Maximilians University Munich (LMU), New York University (NYU), Olin Institute of Living at Hartford Hospital (Olin), Oregon Health and Science University (Oregon), San Diego State University (SDSU), NeuroIMAGE (NI), Stanford University (Stanford), Trinity Centre for Health Sciences (Trinity), University of California, Los Angeles (UCLA), University of Leuven (Leuven), University of Michigan (UMich), University of Pittsburgh School of Medicine (Pitts), University of Utah School of Medicine (Utah), Yale University (Yale)*

	Caltech	CMU	KKI	LMU	NYU	Olin	Oregon	Sdsu
Subjects	38	27	55	57	184	36	28	36
Autism	19	23	22	24	79	20	13	14
Male/Female	32/6	23/4	48/7	48/9	154/30	31/5	24/4	30/6
Age Mean	22.3	25.4	10.4	20	13.9	17.2	10	14.4
Age STD	4.1	4.5	1.4	9.1	5.1	3.2	1.8	1.5

	NI	Stanford	Trinity	UCLA	Leuven	UMich	Pitts	Utah	Yale
Subjects	30	40	49	108	64	145	57	101	56
Autism	15	20	24	62	29	68	30	58	28
Male/Female	24/6	17/3	19/5	52/10	24/5	59/9	25/5	51/7	26/2
Age Mean	29.5	9.5	16.6	12.7	21.4	13.8	17.9	24.5	12.4
Age STD	5.9	1.7	3.0	2.1	2.3	2.7	5.5	3.7	2.9

hold-out set H, contain subjects and their corresponding true labels. Here for each model, the training set (S) is partitioned into five subsets $(S_1, S_2, S_3, S_4, S_5)$ where each subset contains a distribution of class labels, i.e., healthy and ADHD or Autism, in proportion to the whole training set. We also define $S_{-i} = S - S_i$. In each iteration, we use 4/5 of the training set for training and the remaining 1/5 for testing. A different subset is used for testing in each iteration. At each iteration, learner $L(S_{-k})$ where $k \in \{1, 2, 3, 4, 5\}$, learns a classifier C_k from S_{-k}, which is 4/5 of training data. The remaining 1/5 of the training data S_k, is used when computing the test accuracy. Our classifier $C_k(\cdot)$ will output a class label $C_k(x) \in \{$healthy, ADHD (respectively Autism)$\}$ for each subject x and corresponding label y. In general, we define the accuracy of classifier $C(\cdot)$ on set S,

$$\text{acc}_S(C(\cdot)) = \frac{\sum\limits_{(x,y)\in S} I(y = C(x))}{|S|}$$

where $I\{y = C(x)\} \ = \ 1$ if $C(x) \ = \ y$ and 0 otherwise and $|S|$ is total number of subjects in S. The 5-fold cross-validation accuracy is the average of the classification accuracies computed for each of the five cross-validation folds,

$$\text{5CVacc(L, S)} = \frac{1}{5}\sum_{k=1}^{5} acc_{S_k}(C_k)$$

Also, for hold-out set H, the accuracy (for classifier C learned from the training set S) is given by

$$\text{Test Accuracy} = acc_H(C(\cdot))$$

We computed various statistics on the hold-out set, including accuracy, sensitivity, specificity and J-statistics whenever we compared our results to previous results. If we consider label 1 for ADHD-positives (respectively Autism) and label 0 for healthy controls, then

$$\mathrm{Sensitivity}_H(\mathrm{C}(\cdot)) = \frac{\sum_{(x,y)\in H} I\{y=1\}\; I\{C(x)=1\}}{\sum_{y\in H} I\{y=1\}}$$

and

$$\mathrm{Specificity}_H(\mathrm{C}(\cdot)) = \frac{\sum_{(x,y)\in H} I\{y=0\}\; I\{C(x)=0\}}{\sum_{y\in H} I\{y=0\}}$$

and

$$\mathrm{Jstat}_H(\mathrm{C}(\cdot)) = (\mathrm{Sensitivity}_H(\mathrm{C}(\cdot)) + \mathrm{Specificity}_H(\mathrm{C}(\cdot)) - 1)$$

2.3 Preprocessing Pipeline

For preprocessing, we used SPM8 [5] and our own in-house MATLAB code. Our preprocessing involved 6 steps:

1. 6-parameter rigid body motion correction of functional scans
2. Co-registration of functional scans to subject-specific structural scans to guide the spatial normalization step
3. Non-linear spatial normalization (parameter estimation and spatial transformation) of structural images to the MNI T1 template [6]
4. Non-linear spatial normalization of previously co-registered functional image volumes (in step 2) to MNI T1 template using warping parameters computed in the structural image normalization
5. Spatial smoothing of functional image volumes with 8mm full width half maximum (FWHM) Gaussian kernel
6. Z-normalization of each 3D volume intensities for structural and functional image to standardize the intensities of images scanned from different sites.

The details of the pre-processing can be found in Ghiassian et al. [10].

[5] http://www.fil.ion.ucl.ac.uk/spm/software/spm8/

[6] http://imaging.mrc-cbu.cam.ac.uk/imaging/Templates

2.4 Diagnosis from Structural MRI Scans

This section describes in detail the diagnosis algorithm (for ADHD vs. Healthy and Autism vs. Healthy) from MRI scans. We first define some common terms to be used later in this section.

Filters

In image processing, filters are transformations that accentuate certain features within an image. Filters are generally defined on a neighborhood. For example for a 2D image, a filter

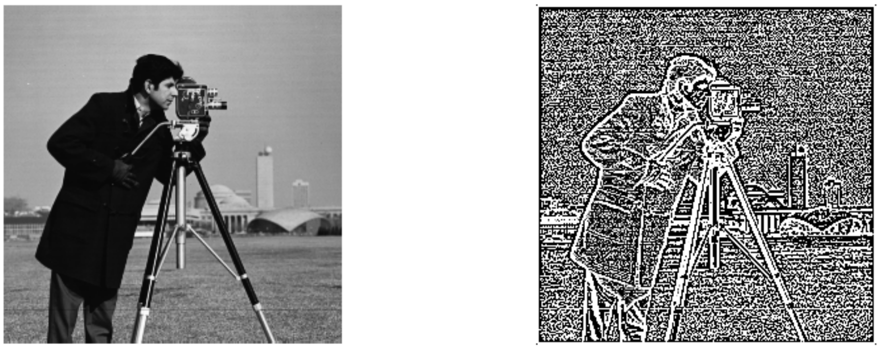
$$h_{contrast} = \begin{bmatrix} 0 & 1 & 0 \\ 1 & -4 & 1 \\ 0 & 1 & 0 \end{bmatrix}$$

is defined on 3×3 neighborhood for each pixel.

Convolution

Convolution is a mathematical operation on two functions $I(x, y)$ and $h(u, v)$, producing a third function that is typically viewed as a modified version of $I(x, y)$, giving the overlap between the two functions. In case of 2-dimensional convolution, I is the image and h is the filter that is convolved with the image. The result of the convolution can be interpreted as the similarity measures between each pixel of the image and the filter. For 2D images, this operation at pixel (x, y) is defined by $g(x, y) = \sum_{(u,v) \in V} I(x - u, y - v)\ h(u, v)$ where V is the neighborhood where filter h is defined as above.

The effect of convolving the image in Fig 2.1a with the filter after convolution is shown in Fig 2.1b. We see the filter h enhances the contrast in the image.



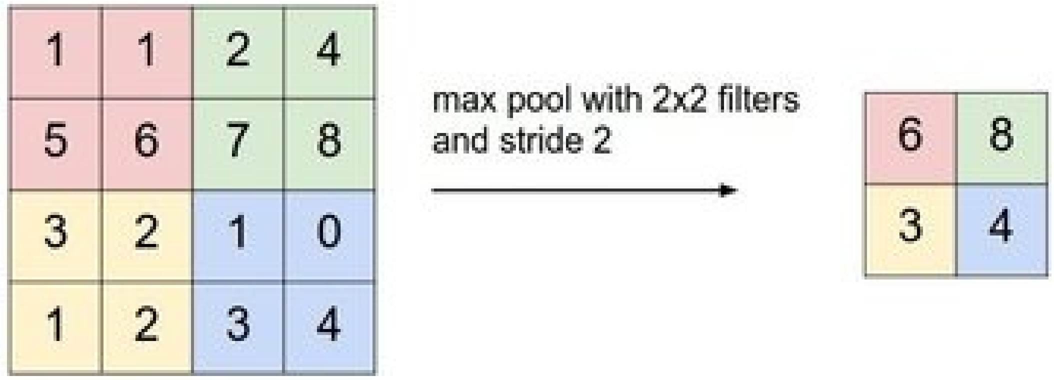
(a) Actual image

(b) After convolution

Figure 2.1: Simple example showing effect of convolution and filtering with $h_{contrast}$ on an image

Max-Pooling

To aggregate statistics of features extracted after convolution (which is an integral part of convolutional neural network described below), we compute the maximum value of a particular feature over a region of the convolved image. We divide our convolved features (in our example before, they are pixels of the convolved image) into disjoint regions, and take the maximum feature value over each region to obtain the pooled convolved features. These summary statistics are much lower in dimension (compared to using all of the pixels from the convolved image) which means they might also improve results (less over-fitting). The aggregation operation is called max pooling. The max-pooling operation is illustrated in Fig. 2.2. The same concept applies to 3D, where instead of 2D filters, 3D filters are used and max-pooling is done in 3D regions.



1	1	2	4
5	6	7	8
3	2	1	0
1	2	3	4

6	8
3	4

Figure 2.2: Max pooling operation. Left: Image after convolution step. Right: After maxpooling step. Stride is distance between two max-pooling regions.

Convolution and maxpooling steps are used in our experiment (**"Method 2"**) whenever we extract features from 3D MRI scans.

2.4.1 Method 1 (Using off-the-shelf Features from Convolutional Neural Network)

There are 2D filters available from other research groups [37, 38], and it is desirable to use those existing 2D filters because they have shown good performance on other classification tasks [39]. However, using those 2D filters with 3D MRI data presents challenges. For example, the 3D geometric relations are not incorporated whenever we use the 2D filters as feature extractor. Also, psychiatric diseases like ADHD/Autism have been associated with functional dysfunction of brain regions. In this case, fMRI is useful to study functional impairment of brain regions. Using only 3D structural MRI scans we are not using the time domain information from fMRI. To address the first challenge, we extract the features in the following way: Use 2D filter extractor from each 2D axial slice and combine the features (described below). To address the second challenge, we devise an algorithm in Section 2.5. This algorithm decomposes fMRI scans into common spatial components and corresponding time courses and uses the time courses as features for disease prediction.

We used *filters* learned from a publicly available trained convolutional neural network (`Overfeat` [7] [38]) that was trained on the *ImageNet*[8] dataset. `Overfeat` has two different learned filter banks described as a) *fast*, vs. b) *accurate*. Among these two, the *accurate* model has more layers and more learned filters. We followed Razavian et al. [39] and used the *accurate* model.

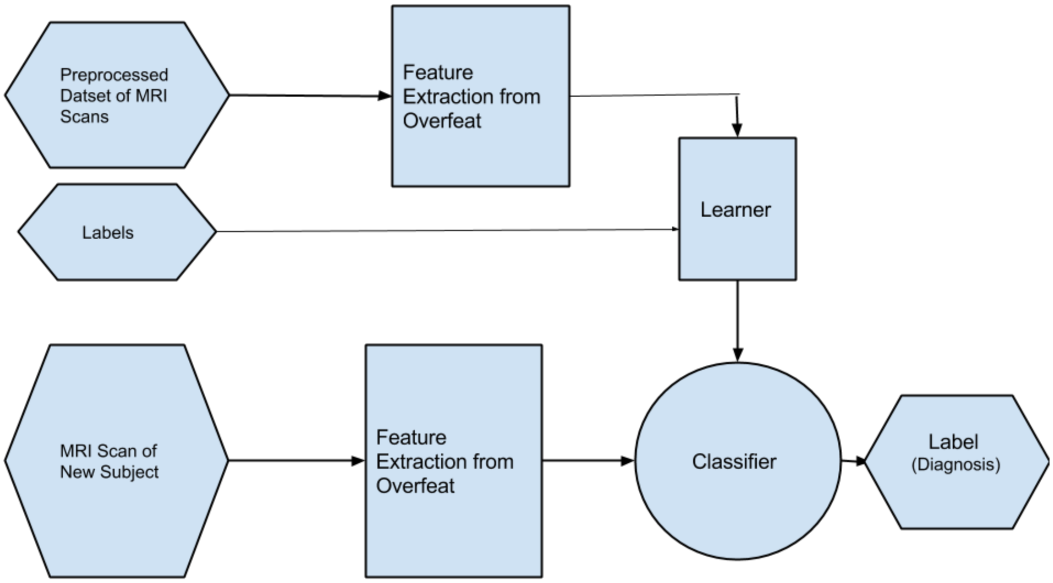

Preprocessed Datset of MRI Scans
Feature Extraction from Overfeat
Labels
Learner
MRI Scan of New Subject
Feature Extraction from Overfeat
Classifier
Label (Diagnosis)

Figure 2.3: Input, algorithm pipeline and output of the learning and performance task for **Method 1**. In this case, each block at the left of an arrow is input to the block at the right of an arrow.

The *feature extraction* block consists of 8 layers of learned filters. At each layer, the output from the previous layer is either convolved with a set of filters (in the case of convolution layers) or else submitted to the max-pooling operator (in the case of max-pooling layers). Filter size for the layers varied from 7 × 7 to 3 × 3. The input of the feature extractor is a 2D image of size 221 × 221. For our experiments, each of the 2D axial slices (of size 79 × 95) from an MRI scan was upsampled with linear interpolation using the **upsample** function in the python **multirate** toolbox to the size 221 × 221 and fed to the

[7] `http://cilvr.nyu.edu/doku.php?id=software:overfeat:start`

[8] `http://www.image-net.org/`

feature extractor.

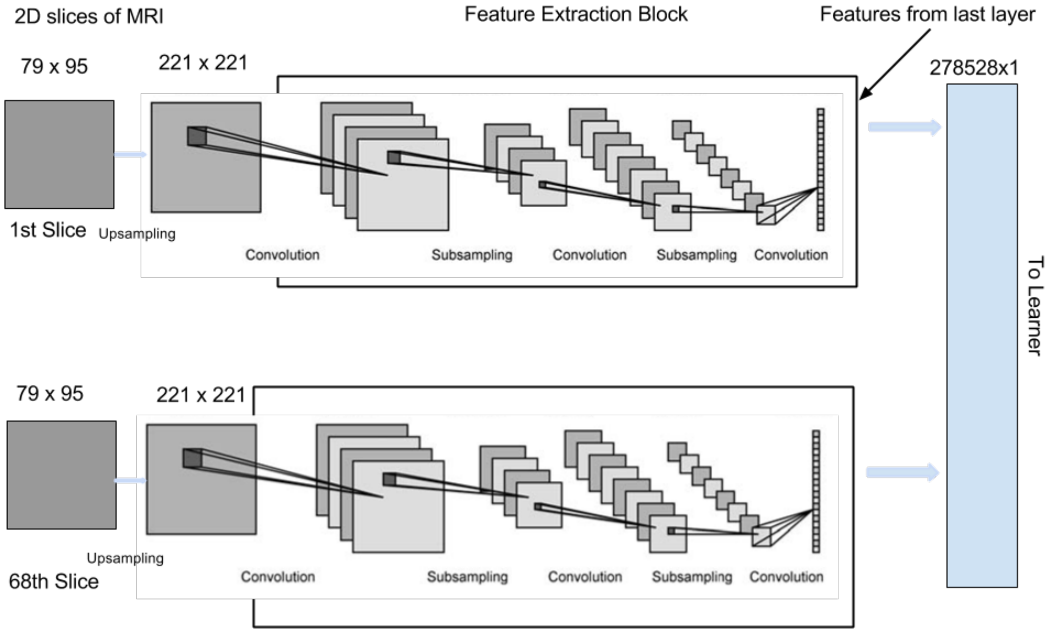


Figure 2.4: Convolutional neural network feature extraction [35]

The feature extraction step is shown in Fig. 2.4. The last layer feature values of the convolutional neural network (before the learning block, 4096 values for an image) were stacked to form one feature vector for one MRI image. This produces 4096 × 68 (where 68 is the number of axial slices) feature values for one MRI scan. These stacked feature values were used to train a linear support vector machine, that is optimizing the following optimization function, for the training data $(\mathbf{x_i}, y_i)$,

$$\min_{\mathbf{w}} \frac{1}{2}\|\mathbf{w}\|^2 + C\sum_i \max(1 - y_i \mathbf{w}^{\mathbf{T}}\mathbf{x_i}, 0) \tag{2.1}$$

where $\mathbf{w}$ is a normal vector to the hyperplane that divides $\mathbf{x_i}$s into different classes. Here $\mathbf{w}$ is a vector of weights learned by the standard SVM learner. This optimization was solved using the kernel trick within the libsvm package [9]. The hyperparameters are learned using 5-fold internal cross validation in each fold.

[9] http://www.csie.ntu.edu.tw/~cjlin/libsvm/

2.4.2 Method 2 (Using a Single Layer Unsupervised Convolutional Neural Net)

In this experiment we use a very simple learning algorithm: just a one-layer unsupervised convolutional network followed by a simple linear support vector machine learner to learn the local statistics from the ADHD-200/ABIDE structural data. After establishing the learning framework, our system used cross-validation to choose the hyperparameters Fig. 2.5.

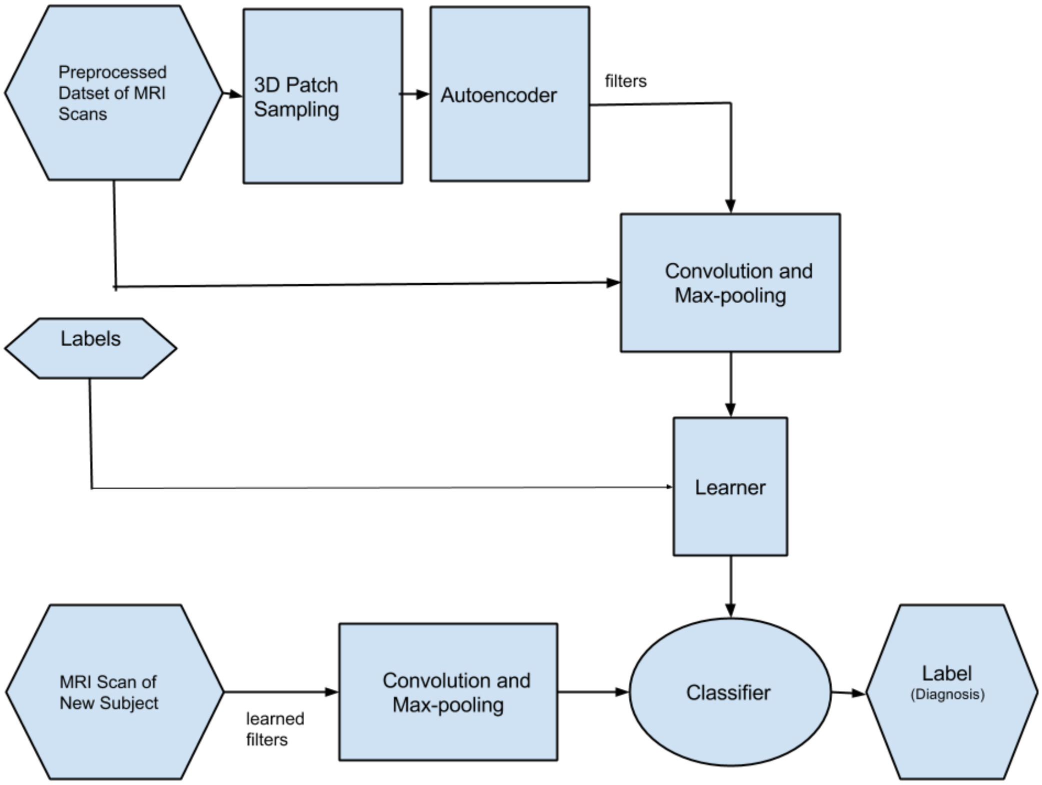


Figure 2.5: Input, algorithm pipeline and output of the learning and performance task for **Method 2**

Unsupervised Feature Learning using Sparse Autoencoder

An autoencoder [33] is a network where the output is the same as input. In our model, the network parameters minimize reconstruction error using the back-propagation algorithm. The network parameters give rise to learned filters. A simple autoencoder is shown in Fig. 2.6,

To avoid the nodes (points to which the inputs are connected) in the hidden layer learning same or redundant filters, sparsity is enforced. An autoencoder is sparse when most of the nodes in the hidden layer are zero given any input. The sparsity of an autoencoder encourages the network to learn different transformations in each of the nodes in hidden layer [34]. This improves robustness in the learned filters – i.e. each filter learned will be

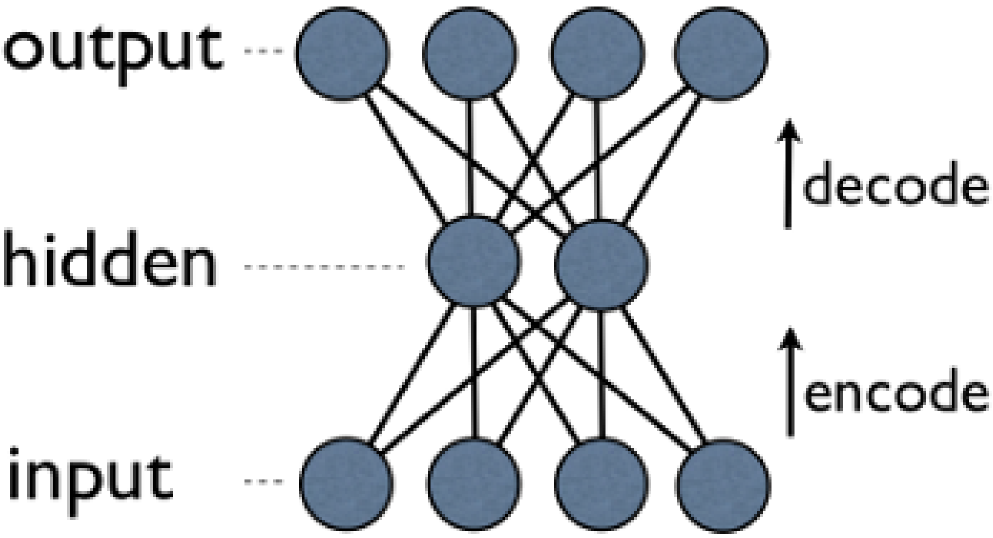


Figure 2.6: A simple autoencoder reference [56]

different and will specify a particular characteristic of the input. Formally the network tries to optimize the following optimization function,

$$J(W, \mathbf{b}) = \frac{1}{2|D|} \sum_{\mathbf{x} \in D} L(\mathbf{x}, \hat{\mathbf{x}}) + \beta \sum_{j=1}^{k} KL(\rho \| \hat{\rho_j}) + \lambda \|W\|_2$$

$$\hat{\mathbf{x}} = \sigma(W_2 \mathbf{h} + \mathbf{b_2})$$

$$\mathbf{h} = \sigma(W_1 \mathbf{x} + \mathbf{b_1})$$

where $D \in R^{m \times n}$ is the data matrix, each data point $\mathbf{x} \in R^n$, also $\mathbf{h} \in R^k$ is the hidden representation of the data, $\hat{\mathbf{x}} \in R^n$ is reconstructed data, $L(\mathbf{a}, \mathbf{b}) = \sum_i (a_i - b_i)^2$ is squared loss error, $\sigma(s) = 1/(1 + e^{-s})$ is sigmoid function, ρ is sparsity parameter, $\hat{\rho}$ is average activation, and $KL(\rho \| \hat{\rho_j}) = \rho \log(\frac{\rho}{\hat{\rho_j}}) + (1 - \rho) \log(\frac{1-\rho}{1-\hat{\rho_j}})$ is Kullback-Leibler divergence, ($W = [W_1, W_2]$ and $\mathbf{b} = [\mathbf{b_1}, \mathbf{b_2}]$) are the parameters to learn and k is the number of nodes in the hidden layer. The weights of the connections learned at each node corresponds to one filter. The hyperparameters are learned using 5-fold cross validation. The optimization was solved using L-BFGS package [10].

Convolutional Network

Once the filters (total number k) are learned from sparse autoencoder, the each filter is convolved with the actual image to produce feature maps. An example of the features learned from natural images is shown in Fig 2.7.

The convolved features produced are highly non-linear. This comes from the sigmoid function in the cost [35]. In order to reduce the dimension of the features, a subsequent

[10] `http://users.iems.northwestern.edu/~nocedal/lbfgsb.html`

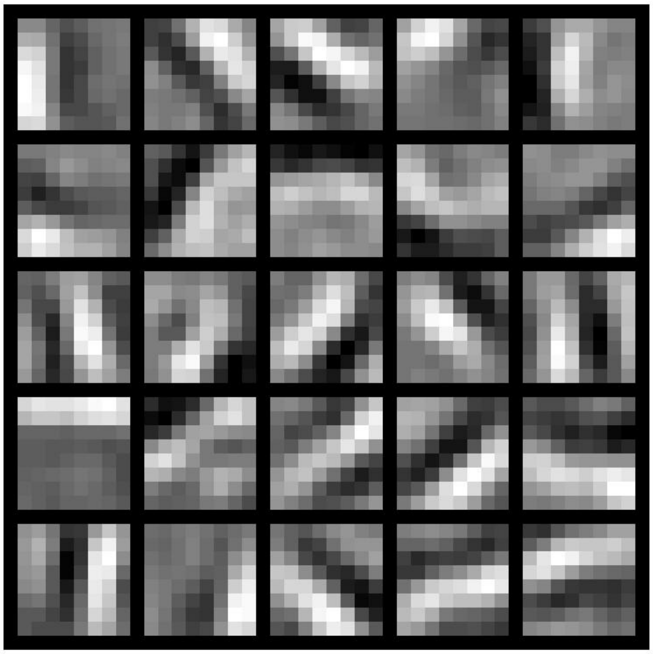


Figure 2.7: 25 filters learned from sparse auto encoder using natural image patches

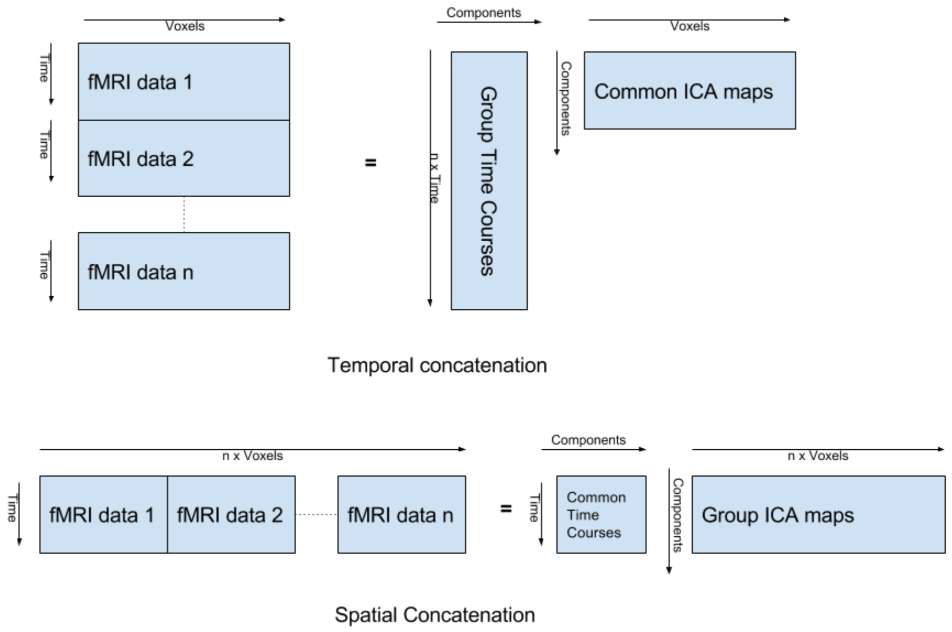
layer after the convolution performs local pooling thus reducing the resolution of the feature maps. This introduces translation invariance of the features in the max-pooling region. Max pooling was used in the experiment using a $5 \times 5 \times 5$ non-overlapping bounding box. This reduces the feature dimensions by a factor of 125. This step also reduces sensitivity of the feature map to various distortions. These extracted features were later used for learning and prediction.

Learner

The features were used to train a linear support vector machine where for the training data $\{(\mathbf{x_i}, y_i)\}$, we perform the optimization as described in Eqn 2.1.

2.5 Diagnosis from fMRI Scans

This section describes the diagnosis algorithm for ADHD vs. Healthy and Autism vs. Healthy prediction from fMRI scans using blind source separation. Source separation techniques can identify functionally connected networks by estimating spatially independent patterns from their linearly mixed fMRI signal [46]. However source separation for different individuals can produce different spatial patterns (also known as spatial maps) for different subjects and hence temporal or spatial concatenation becomes necessary before applying any source separation method. Each source separation method decomposes fMRI scan into spatial components or maps with associated time courses. This means that each voxel in the spatial map will have common time course. We use temporal concatenation approach that allows for unique time courses (TCs) for each subject, but assumes common spatial maps (SMs) across all subjects whereas the spatial concatenation approach (not discussed here) allows for unique SMs but assumes common TCs. Although they are really just two different approaches for organizing the data (spatial vs temporal) as shown in Fig. 2.8, temporal concatenation appears to work better for fMRI data [44] most likely because the subject-to-subject temporal variations are much larger than the variation in the spatial maps at conventional field strengths of 3T and below [43, 46].



Temporal concatenation

Spatial Concatenation

Figure 2.8: Temporal vs. spatial concatenation

fMRI scans have a large number of voxels over a number of time points. Combining all the fMRI scans from all subjects for source separation becomes computationally intensive and intractable. Hence before the concatenation, dimensionality reduction of the data becomes necessary [1],[2] to capture the subject level variations in the data. This is done using principal component analysis (PCA) on individual subjects. After that our algorithm concatenates the fMRI scans and source separation is done [46]. Previous studies([1, 52]) have explored brain regions that are strongly temporally coherent (which means they are co-activated during rest) using source separation techniques like PCA and independent component analysis (ICA). PCA separates the fMRI brain scan into uncorrelated spatial maps or sources based on variations in the time whereas ICA decomposes the brain fMRI scans into spatially independent components (sources or spatial maps) and their corresponding time courses. It assumes that the spatial maps have constant higher order statistics [1, 2, 4]. For example, one source may be a random process with probability density function $\frac{\lambda}{2}\exp(-\lambda|x|)$ (the probability of seeing a voxel with intensity x) with constant parameter λ. On the other hand, a variable parameter will be dependent on the location of the brain region. An example of probability density function with a variable parameter $\lambda(r)$ (r is location of the voxel in the scan) may be denoted as be $\frac{\lambda(r)}{2}\exp(-\lambda(r)|x|)$. Here λ is parameterized by r. Here we introduce *k-PCA* and *NSD* for fMRI source separation. In all the cases, we denote the fMRI scan of the i^{th} subject as $X^i_{[T\times V]}$ and view $X^i_{[T\times V]} \approx A_{i[T\times K]} \times S_{[K\times V]}$ where the rows of S are estimated spatial maps and the columns of A_i are corresponding estimated time courses. Note that the S matrix is the same for all users, but the A matrix varies from patient to patient.

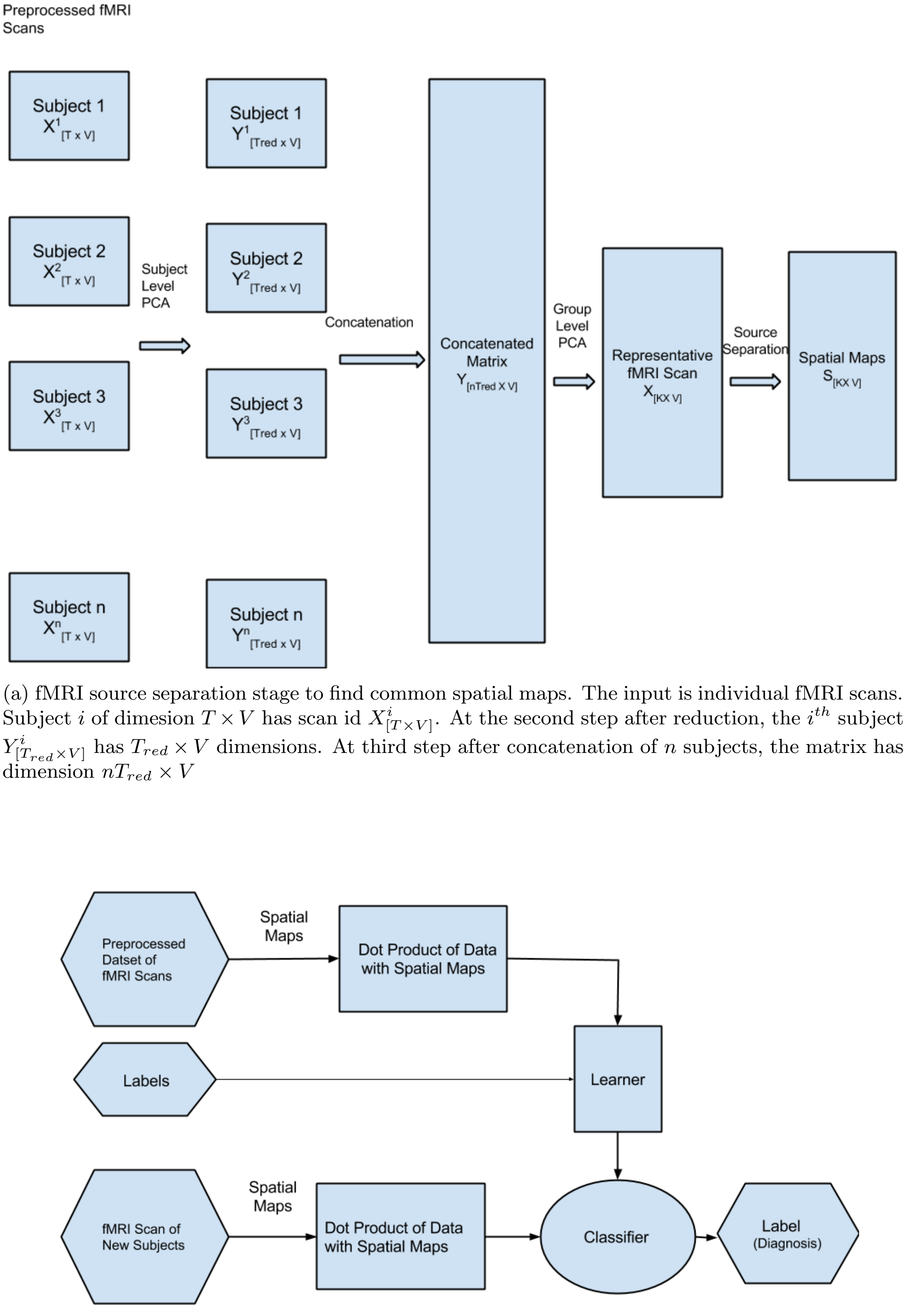
The common use of ICA based source separation in fMRI applies the infomax [51] principle for separating components that decompose the brain into spatially independent maps. However, the theoretical derivation of ICA requires the following assumptions to hold: i) the spatial maps should have non-gaussian distribution, and ii) each voxel time course in the fMRI scan should be independent and identically distributed (IID). We relax these assumptions and experiment with four different source separation techniques to find a good representation of fMRI scans using common bases. The input to this algorithm is preprocessed fMRI scans.

2.5.1 Method 3 (Using fMRI Source Separation Models for Prediction of Psychiatric Diseases)

Our algorithm pipeline is given in Fig. 2.9. For source separation we compare each of the four separation methods and compare the results in the next chapter.

Dimensionality Reduction

In order to reduce the computational load on group level analysis, different approaches have been proposed for dimensionality reduction before group level source separation analysis. We follow the standard 2-Step principal component analysis reduction of the data [1],[2]. Our main motivation for using this model is its assumption that there are common spatial sources for each data set where subjects differ based on temporal weights of each source. The 2-Step data reduction captures the subject level variations and group level commonalities.


Preprocessed fMRI Scans
Subject 1 $X^1_{[T \times V]}$
Subject 2 $X^2_{[T \times V]}$
Subject 3 $X^3_{[T \times V]}$
Subject n $X^n_{[T \times V]}$
Subject Level PCA
Subject 1 $Y^1_{[Tred \times V]}$
Subject 2 $Y^2_{[Tred \times V]}$
Subject 3 $Y^3_{[Tred \times V]}$
Subject n $Y^n_{[Tred \times V]}$
Concatenation
Concatenated Matrix $Y_{[nTred \times V]}$
Group Level PCA
Representative fMRI Scan $X_{[KX V]}$
Source Separation
Spatial Maps $S_{[KX V]}$
Preprocessed Datset of fMRI Scans
Spatial Maps
Dot Product of Data with Spatial Maps
Labels
Learner
fMRI Scan of New Subjects
Spatial Maps
Dot Product of Data with Spatial Maps
Classifier
Label (Diagnosis)

(a) fMRI source separation stage to find common spatial maps. The input is individual fMRI scans. Subject i of dimesion $T \times V$ has scan id $X^i_{[T \times V]}$. At the second step after reduction, the i^{th} subject $Y^i_{[T_{red} \times V]}$ has $T_{red} \times V$ dimensions. At third step after concatenation of n subjects, the matrix has dimension $nT_{red} \times V$

(b) Input, algorithm pipeline and output of the learning and performance task for **Method 3**

Figure 2.9: Our learning system in two stages. First stage finds common spatial maps. The second stage develops the classifier responsible for diagnosing new subjects, which is then used in testing.

Theory of Principal Component Analysis (PCA): For any data matrix $X_{[p \times q]}$, where p is number of features and q number of examples, we can define a covariance matrix of the features (assuming X is zero centered),

$$\Sigma = XX^T \tag{2.2}$$

Then $\Sigma \mathbf{e} = \lambda \mathbf{e}$ holds for each eigenvalue/eigenvector pair $(\lambda, \mathbf{e})$, where eigenvector $\mathbf{e} \in \Re^{\mathbf{p}}$ has the corresponding eigenvalue $\lambda \in \Re$. Since Σ is symmetric, it will always have non-negative eigenvalues $\lambda \in \Re$. We can sort eigenvalues of Σ in descending order, i.e. $\lambda_i \geq \lambda_{i+1}$. There can be at most p such eigenvalue/eigenvector pairs (assuming $p \leq q$). The eigenvectors of a matrix are orthogonal to every other eigenvector of this matrix. The i^{th} principal component is the data matrix projected onto eigenvector $\mathbf{e_i}$.

In order to apply source separation for the whole dataset, we apply a data reduction step following Calhoun et al. [1]. This data reduction procedure uses 2-Step principal component analysis and is implemented in the GIFT package [11]. We briefly describe the ideas involving the data reduction stage. Suppose we have fMRI scan X^i matrices for $i = 1, .., n$ subjects. We then reduce the $T \times V$ data matrix X^i from each subject, to a $T_{red} \times V$ matrix by selecting T_{red} largest eigenvalues capturing 99% of the variance using PCA. Next, the reduced data-matrices from each subject are concatenated to form the group level data matrix of size $nT_{red} \times V$. The concatenated dataset is then passed through a second group level PCA to select K dimensions. After this 2-Step data reduction, we get an aggregate data matrix of size $K \times V$ over all subjects. Source separation is performed on this data matrix. This data matrix, denoted as $X_{[K \times V]}$, is designated to represent the corresponding dataset.

Here we consider the model, $X^i_{[T \times V]} = A_{i[T \times K]} S_{[K \times V]}$ following [4]. The model assumes that the response of each fMRI voxel at time t is a weighted linear combination of specific sources common across subjects. But the weighting of these sources (time courses for each source) for different subjects will be different.

In the first level PCA, the reduced matrix for subject i is $Y^i = U_i X^i$ where U_i is the $T_{red} \times T$ reduction matrix. Then we concatenate Y^i's to get Y of size $nT_{red} \times V$. In the second step, the reduced matrix $X = FY$ where F is a $K \times nT_{red}$ reduction matrix. If we divide the F into n sub-blocks F_i, each of size $K \times T_{red}$, then F_i corresponds to reduction matrix for subject i. This is shown pictorially in Figures 2.9a and 2.10. Hence for subject i,

$$X^i = U_i^- F_i^- X = U_i^- F_i^- AS = A_i S \tag{2.3}$$

where $A_i = U_i^- F_i^- A$, which is different for each patient. Here U_i^- and F_i^- are pseudo-inverses of U_i and F_i respectively. X^i is represented by A_i and S pictorially in Fig. 2.10

[11] http://mialab.mrn.org/software/gift/

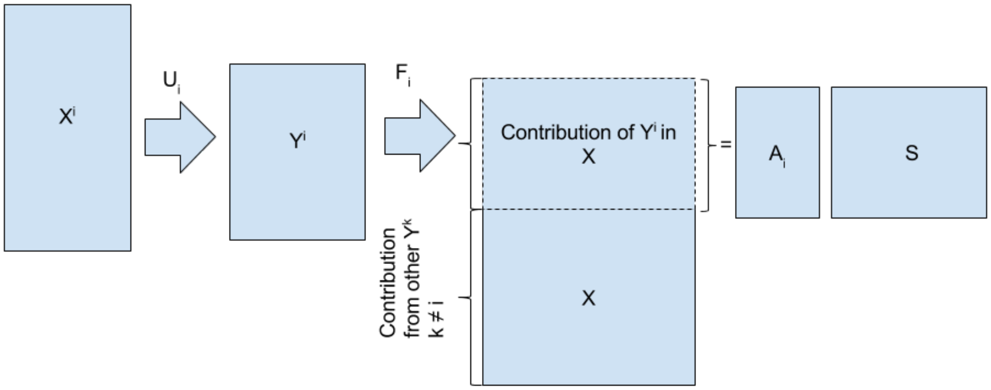


Figure 2.10: Pictorial Representation of X^i by A_i and S. Here each arrow implies dot product of the matrix before arrow with the matrix on the arrow

Specifically for ADHD or Autism prediction, our hypothesis is: the time courses for the patients will be different from controls for the ADHD/Autism patients. This hypothesis is validated by the 5-fold cross validation and test accuracy on ADHD-200/ABIDE data.

Separation of Spatial Sources

Principal Component Analysis (PCA) for fMRI In this case, after dimensionality reduction from Equation 2.3, the rows of representative scan $X_{[K \times V]}$ are directly used as spatial maps. The projection of i^{th} patient's fMRI scan on these vectors yields the time components (columns of A_i).

Theory of Kernel Principal Component Analysis (k-PCA) For representative fMRI scan $X_{[K \times V]}$, a nonlinear similarity can exist in *inner-product* space H (such that $\phi : X \rightarrow H$).

$$\Sigma_\phi = \phi(X)\ \phi(X)^T \tag{2.4}$$

For *Radial Basis Function kernel* (RBF),

$$\Sigma_{rbf}(i,j) = \exp(-\frac{(x_i - x_j).(x_i - x_j)}{2\sigma^2}) \tag{2.5}$$

where x_i and x_j are i^{th} and j^{th} column of X respectively. Once the kernel similarity matrix is computed, we get the eigenvalue/eigenvector pair of the kernel matrix. The projection on eigenvectors would be the spatial maps (rows of S) and the dot product of each patient's fMRI scan on these maps are the time components (columns of A_i).

Theory of Independent Component Analysis (ICA) Current source separation methods in the fMRI literature mostly focus on separating statistically independent stationary signals over all voxels using ICA. Here we give log-likelihood interpretation for ICA. The process for independent component analysis is shown in Fig 2.11.

For representative fMRI scan $X_{[K \times V]} \approx A_{[K \times K]} \times S_{[K \times V]}$, we denote each row i of X as x^i (size $1 \times V$) and each column j as x_j (size $K \times 1$). Assume each row x^i consists of V observation of the random variable $\mathbf{x^i}$. We denote each row i in S by s^i and assume row s^i consists of V observations of random variable $\mathbf{s^i}$. We also denote each column j of S as s_j. Now our goal is to find A such $\mathbf{s^i}$ for $i \in \{1, 2, 3...K\}$ are independent of each other.

$$x_j = As_j, \ s_j = Wx_j \tag{2.6}$$

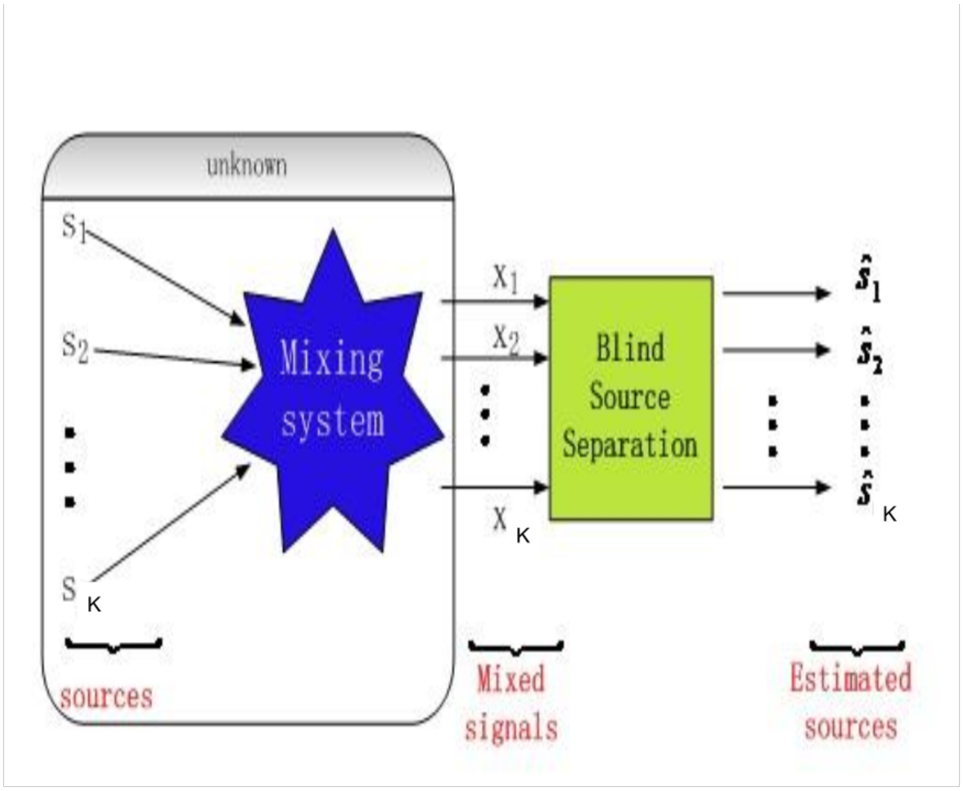


Figure 2.11: Independent component analysis illustration reference [57]

where W is inverse of A. Repeated observations of x_j at different voxels ($j \ = 1, 2..V$ i.e. time series for V different voxels) give us each column of X. To capture log-likelihood formulation of ICA, for the data X, we assume that the distribution of i^{th} source $\mathbf{s^i}$ is given by a density $p_s(\mathbf{s^i})$, and that the joint distribution of the sources are independent and so is given by

$$p(\mathbf{s^1}, \mathbf{s^2}, ..\mathbf{s^K}) = \prod_{i=1}^{K} p_s(\mathbf{s^i}) \tag{2.7}$$

The probability of the received signals for j^{th} observation (x_j) is given by (see independent component analysis[12])

$$P([\mathbf{x^1}, \mathbf{x^2}, ..., \mathbf{x^K}]^{\mathbf{T}} = x_j) \propto \prod_{i=1}^{K} p_s(W_{i,:}^T x_j) \times \det(W) \tag{2.8}$$

Where W^T is transpose of W. The log-likelihood formulation is

$$\max_W \ \big(\sum_{i=1}^{K} \log \ p_s(W_{i,:}^T x_j) + \log \ \det(W)\big) \tag{2.9}$$

Any nongaussian distribution can be assumed for p_s[12]. For the dataset $(X_{[K \times V]})$, the optimization function to maximize is

$$W^* \ = \ \operatorname*{argmax}_W \ \big(\sum_{i=1}^{K}\sum_{j=1}^{V} \log \ p_s(W_{i,:}^T x_j) + \log \ \det(W)\big) \tag{2.10}$$

After estimating W^* on the reduced matrix, we estimate the common spatial sources S. Then, for each patient, the time courses A_i are calculated from Equation 2.3.

Theory of Non-Stationary Spatial Sources Decomposition (NSD) The principle of source separation using independent component analysis is based on the assumption that each source has constant higher statistics − i.e. as mentioned earlier, the probability density of each source is parameterized by a constant value which does not change with location of voxels. Here two neighboring voxels are assumed to be independent. In contrast, our model allows two neighboring voxels to be correlated. During source separation, we model each scan to be a combination of spatial maps (or sources) where each spatial map is non-stationary − i.e. it is taken from a probability density function parameterized by location of voxels. In this case, it can be shown that the separation can be based on multiple decorrelation at different locations [54, 55]. We think these spatial sources should be non-stationary as:

1. One source (which corresponds to one row in S) may not have the same magnitude and variation throughout the whole brain due to in-homogeneous magnetic susceptibility that depends on the location of voxels in the scan. Commonly used source separation models require that the sources have same variation for the whole brain scan.

[12] http://cs229.stanford.edu/notes/cs229-notes11.pdf

2. The strength and variability within a particular source (which corresponds to one row in S) depends on the brain tissue type in a particular brain region. For example, activation values in grey matter and white matter (which depends on the amount of oxygenated blood flow in that tissue) would be different.

In order to develop the theory for nonstationary source decomposition, we note, the representative fMRI scan $X_{[K\times V]} \approx A_{[K\times K]}S_{[K\times V]}$, where S is an $K \times V$ matrix of source components, where each row k (size $1 \times V$) provides contribution of voxel co-ordinates to k^{th} source. Each spatial activation refers to one row in S. Column k in matrix A will have corresponding time courses for k^{th} spatial component.

In mathematical terms, suppose we have K independent spatial maps (corresponding to each row in S) and V observations, each corresponding to a 3D brain scan at a time point. Then, we can formulate the covariance matrix at location r as $R_x(r) = \langle\, x(r)\, x(r)^T\,\rangle = A\, D_s(r)\; A^T$ where R_x, $D_s(r)$ are of size $K\times K$. Further we let $x(r) = X(:, N(r))$ where $N(r)$ represents any suitably chosen region around position r for which the signals are assumed to have same higher order statistics. For our experiments, we have chosen $N(r)$ to be a $4 \times 4 \times 4$ bounding box (e.g., if the point r = [200, 100, 50], then this box is defined by corners [199, 99, 49] and [202, 102, 52]). Increasing the bounding box severely degraded the performance of the model (5-fold cross validation accuracy described in Section 2.2).

Assuming the sources are spatially non-stationary, and following [7], [8],

$$x(r) \approx A\, s(r)$$

$$R_x(r) = \ \langle x(r)\, x(r)^T\rangle \ = A\ \langle s(r)\, s(r)^T\rangle\ A^T = A\ D_s(r)\ A^T$$

However, as we do not have a perfect estimate for $R_x(r)$, we estimate the covariance matrix $R_x(r)$ for some spatial interval. We denote the sample estimates as $R_x^{est}(r)$. The measurement error

$$E(r) = R_x^{est}(r) - AD_s(r)A^T$$

Suppose we have N samples of $R_x^{est}(r)$ for $\{r \ \in\ r_1, r_2, .., r_N\}$, then we can estimate the parameters by

$$A^{est}, D_s^{est}(r_1), D_s^{est}(r_2), ..D_s^{est}(r_N) = \underset{A, D_s(r_1),...,D_s(r_N)}{\operatorname{argmin}} \sum_{k=1}^{N} \|E(r_k)\|^2$$

This is with high confidence, accurate for large N. Now, the source components can be estimated as

$$s^{est} = \underset{s}{\operatorname{argmin}}\, \|x - A^{est}s\|^2$$

Then, for each patient the time courses A_i is calculated from Eqn. 2.3. This method decorrelates the spatial sources at different regions of the brain which is desirable for the reasons described before.

Support Vector Machine Classifier The extracted time courses for each component yields a total number of $K \times T$ features. They are used as input to an support vector machine (svm) learner with radial basis function kernel (with $\gamma \in V$ where V is $[0.1 : 102.4]$ with $V(i) = 2 \times V(i-1)$; which is a standard practice) to produce a predictor, which can then be used to predict the class of a novel instance.

2.6 Diagnosis from Multimodal Features

2.6.1 Method 4 (Multi-modal Features for Prediction of Psychiatric Diseases)

For n subjects, suppose $X^{mri}_{n \times f1}$ is feature matrix from the one-layer convolutional network and $X^{fmri}_{n \times f2}$ is feature matrix from nonstationary ICA.

The combined feature matrix is $X^{combined}_{n \times f}$; where $f = f1 + f2$. This matrix was then divided into training set and holdout set by instances as in Section 2.1.

Both these sets of feature values were used to train a linear support vector machine where for the training data $\{(\mathbf{x_i}, y_i)\}$, we perform the optimization as described in Eqn 2.1.

Chapter 3

Results

3.1 Method 1 (Using off-the-shelf Features from Convolutional Neural Network)

The 5-fold cross validation accuracy on the ADHD-200 data training set was 0.5986 vs. a baseline of 0.50. The hold-out set accuracy was 0.6023. The sensitivity, specificity and Jstat were 0.3156, 0.7655 and 0.0811 respectively. For the ABIDE dataset, the 5-fold cross validation accuracy on the training set was 0.5237. The test set accuaracy was 0.4951 (less than chance accuracy). The specificity and sensitivity were 0.5625 and 0.4236. Jstat is not reported here as specificity+sensitivity<1.

3.1.1 Discussion

In this experiment, transfer learning from 2D image recognition is not very helpful. The reason for the lackluster performance may be due to inherent 3D geometric structures of brain and different inherent image statistics of MRI images, which is different from identifying objects within 2D images.

3.2 Method 2 (Using a Single Layer Unsupervised Convolutional Neural Net)

3.2.1 Model Accuracy

The results for 5-fold cross validation and hold-out accuracies are shown in Tables 3.1 and 3.2. The best 5-fold cross-validation accuracy achievable for ADHD-200 data is 0.6346. For the hold-out data it is 0.6257. The sensitivity, specificity and Jstat of the classifier on the hold-out set is given by 0.4195, 0.8421 and 0.2616 respectively.

The best 5-fold cross-validation accuracy achievable for ABIDE data is 0.6137. For the hold-out data, it is 0.6173. The sensitivity, specificity and Jstat of the classifier on the hold-out set is given by 0.4896, 0.7296 and 0.2092 respectively.

Table 3.1: 5-fold cross-validation and hold-out results for ADHD classification using features from structural images

Number of Feature Maps	5-Fold CV Accuracy	Hold-out Accuracy
75	0.6201	_
100	**0.6346**	**0.6257**
125	0.6254	_
150	0.6129	_
175	0.6111	_
200	0.6290	_

Table 3.2: 5-fold cross-validation and hold-out results for Autism classification using features from structural images

Number of Feature Maps	5-Fold CV Accuracy	Hold-out Accuracy
75	0.5974	_
100	0.5962	_
125	0.5916	_
150	0.5987	_
175	**0.6137**	0.6173
200	0.6113	_

3.2.2 Discussion

In this experiment, we showed that a simple texture based feature learning method can be useful for the classification and prediction of psychiatric diseases (ADHD/Autism). Especially using the textures that were learned from the data itself, we were able to predict the disease comparably to state-of-the-art accuracies for ADHD [10, 14, 13, 4] and autism [11]. The main reasons for the performance of this simple model are two fold.

- The model incorporated nonlinear transformation in the convolution and max-pooling layer shown in Fig. 2.5 in the form of a sigmoid function. Compared to linear texture model, this can enhance subtle details [32] in a tissue. The median axial slice (34^{th} axial slice as we have 68 axial slices in total for each subject) for one subject from 3D MRI scan is shown in Fig. 3.1. We have also shown the voxel values in the colorbar. Noticeably, most of the voxels have very high values. Figures 3.3a and 3.3b show the effect of the learned filters on median axial slices for one ADHD and one Autism patient respectively. These images were computed by convolving each filter with one ADHD (respectively autism) patient's structural MRI scan. Here, some parts of the brain tissues are prominent (having larger weights) while the other parts insubstantial (having lesser weights) after they are convolved with the learned filters (Fig. 3.2 shows one substantial portion in the tissue after convolution). The arrangement of voxels after the covolution, is very helpful and informative. These features are important

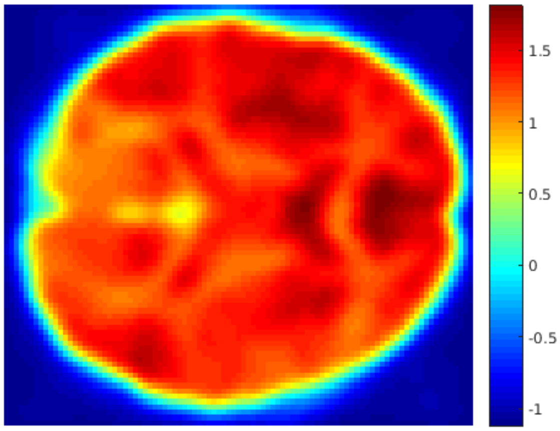


Figure 3.1: Median axial slice from MRI scan

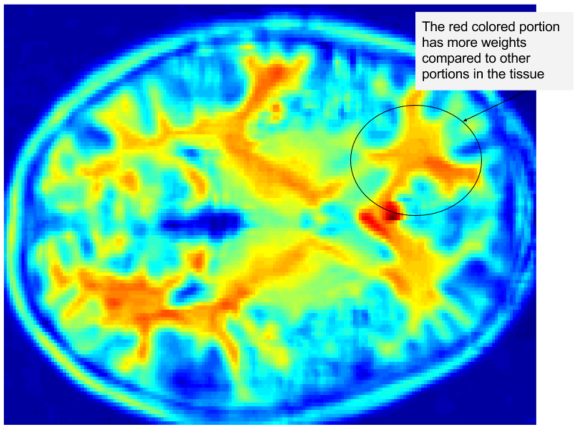
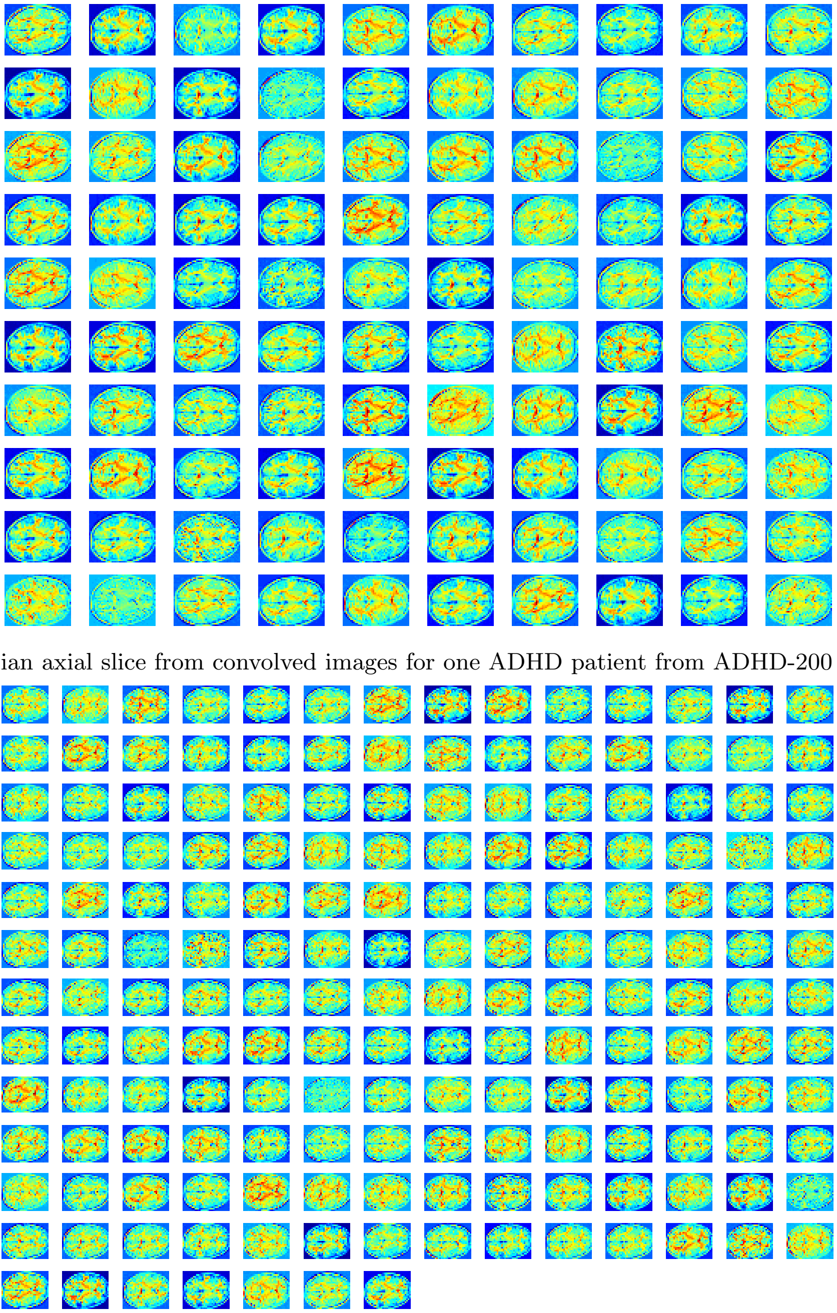

The red colored portion has more weights compared to other portions in the tissue

Figure 3.2: Median axial slice from MRI scan after convolution with one filter

as they may be able to signify loss of brain tissue and inflammation in those regions. Also, the features have good predictive power as can be seen in test accuracies of the model.

- The model captured 3-D texture information in the learned filters shown in Fig. 3.4 and Fig. 3.5. Here we have shown five 5 × 5 2-D axial slices of 5 × 5 × 5 filters for visualization. The filters mainly capture different orientations of edges, blobs and spacial arrangements of voxels. These filters learned are different from the filters learned from natural images shown in Fig. 2.7. These filters learned from MRI data are informative as can be seen from the test set accuracy of the model. Hence learning these domain specific filters helped to improve the prediction accuracy of the model.



(a) Median axial slice from convolved images for one ADHD patient from ADHD-200 dataset

(b) Median axial slice from convolved images for one Autism patient from ABIDE dataset

Figure 3.3: Effect of learned filters on median MRI scan

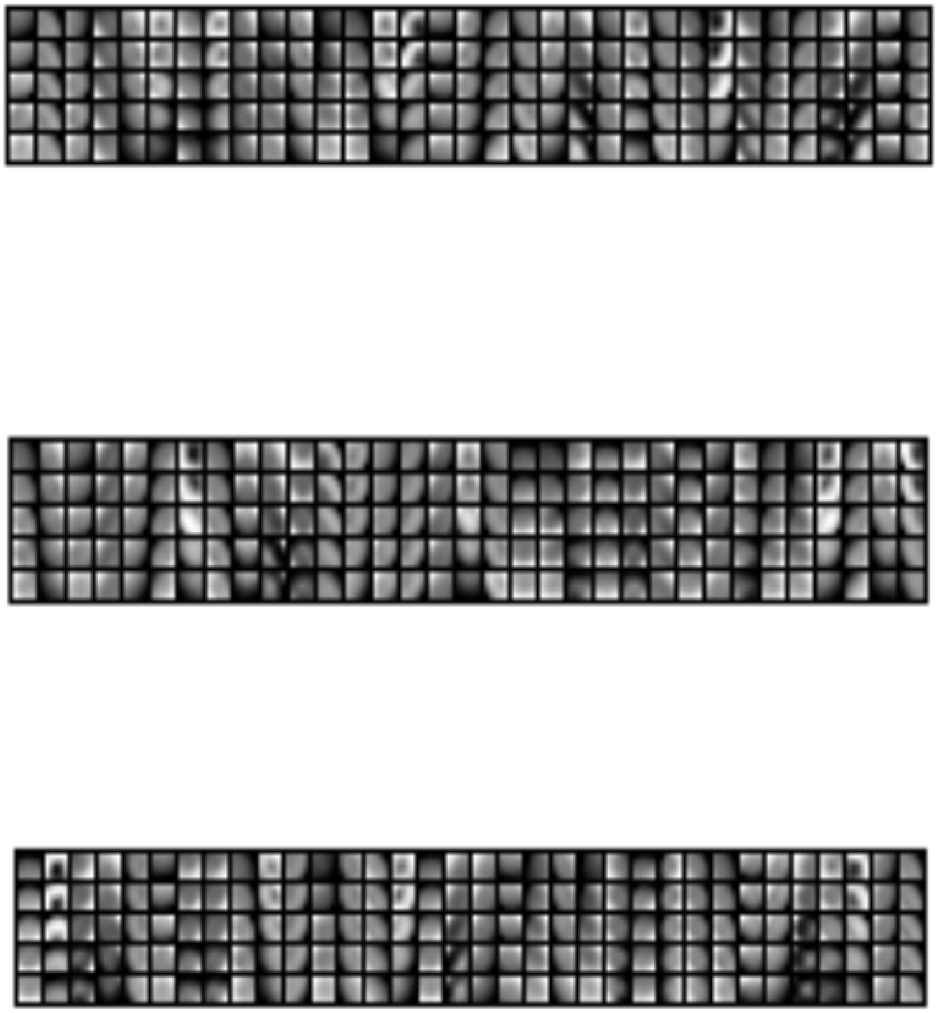


Figure 3.4: 100 bases for ADHD diagnosis. Every column shows the five slices from one filter where each slice is of size 5 × 5. These filters learned are different from the filters learned from natural images shown in Fig. 2.7

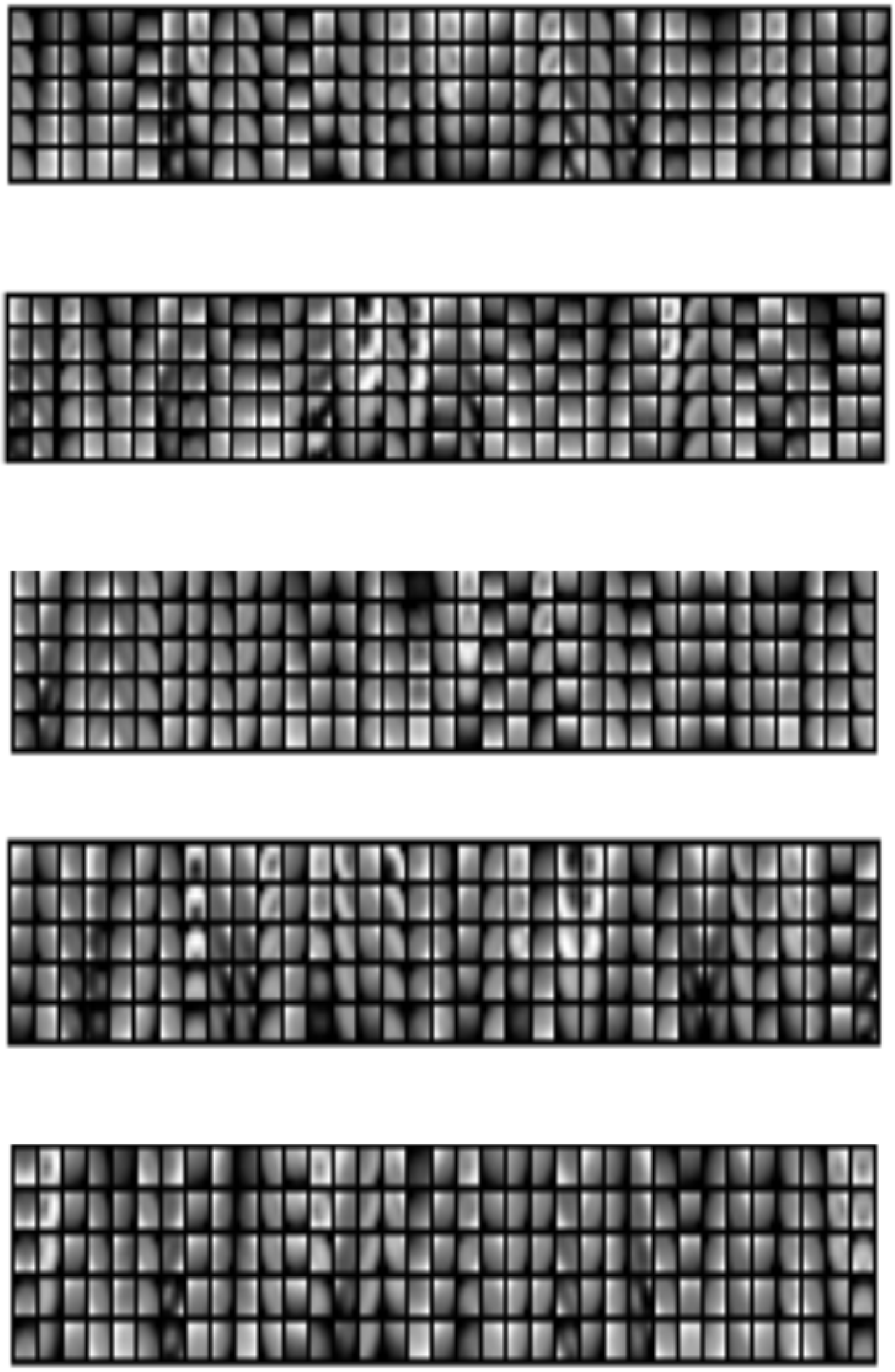


Figure 3.5: 175 bases For Autism diagnosis. Every Column shows the five slices from one filter where each slice is of size 5 × 5. These filters learned are different from the filters learned from natural images shown in Fig. 2.7

3.3 Method 3 (Using fMRI Source Separation Models for Prediction of Psychiatric Diseases)

3.3.1 Experiments Using ADHD-200 Data

Model 5-Fold Cross Validation Accuracy

The results showing our model accuracies on training data (baseline 0.50) are given in Table 3.3 which shows that using 45 independent components provide the best classification accuracy with significant statistical significance (p=1.827e-12) vs. baseline.

Summary of ADHD-200 5-fold cross validation accuracy is given in Fig. 3.6.

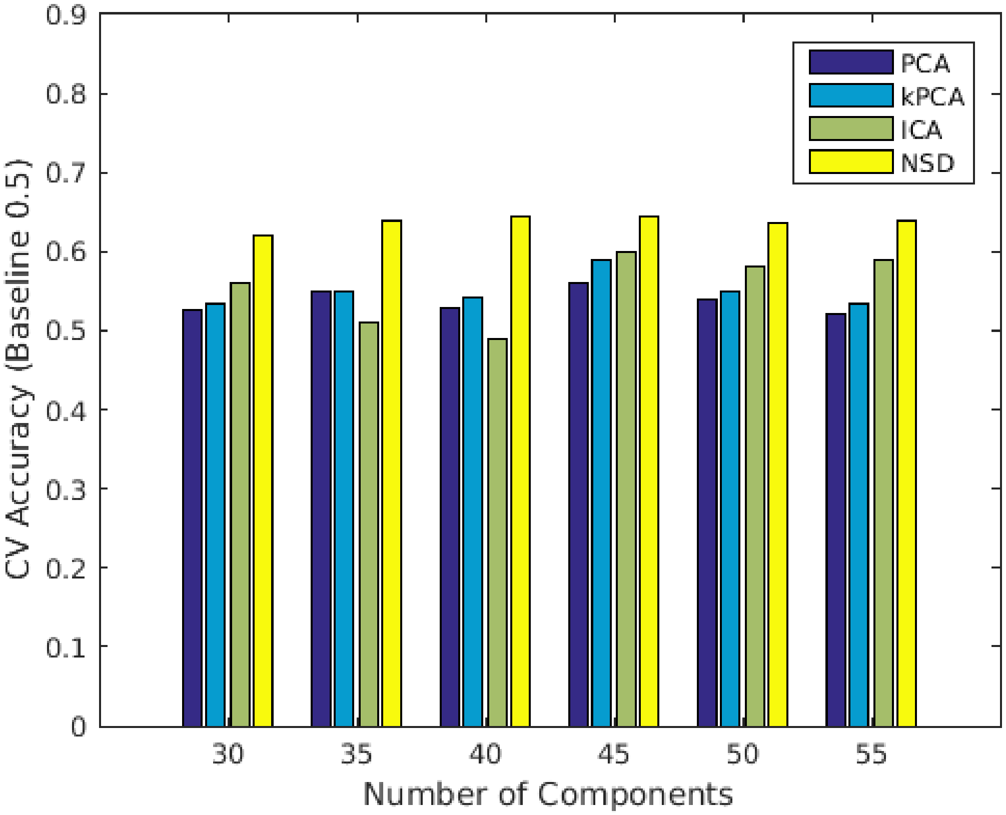


Figure 3.6: 5-fold cross validation results of ADHD classification

Model Hold-out Test Accuracy

The test accuracy was **0.6491** on the hold-out data (baseline 0.5497). To our knowledge, this is the best test-set accuracy achieved on ADHD-200 test data using only fMRI scans. The test-set result is also statistically significant (p=0.0033) vs. the baseline. The specificity, sensitivity and Jstat are given by 0.8191, 0.4416 and 0.2607 respectively. A comparison of our results to previously published best performing algorithms for ADHD-200 competition

Table 3.3: 5-fold cross-validation results for ADHD classification using different number of independent components

Number of ICs (SVM Gamma)	5-Fold CV Accuracy	5-Fold CV STD
30(6.4)	0.6216	0.0399
35(3.2)	0.6378	0.0374
40(1.6)	0.6432	0.0390
45(3.2)	**0.6450**	0.0291
50(3.2)	0.6360	0.0261
55(3.2)	0.6378	0.0075

Table 3.4: Hold-out test results for ADHD classification using only imaging data

Algorithm	Accuracy	Specificity	Sensitivity	J-Statistics
Our Result	**0.6491**	0.8191	**0.4416**	**0.2607**
Eloyan et al [3]	0.610	0.94	0.21	0.15
Dai et al [13]	0.6150	0.7766	0.4133	0.1833
Sidhu et al [14]	0.614	_	_	_
Ghiassian et al [10]	0.6260	_	_	_

is shown in Table 3.4.

Spatial Maps from ADHD-200 data

The spatial maps from non-stationary source decomposition are shown in Fig. 3.7 and Fig. 3.8. For each spatial map, 9 axial slices are shown. IC1 is an artifact as all the voxels have very high values in this spatial map. Because all the voxels in the brain are equally important in this component, it is shared by all the voxels. It has been shown that noise like motion, breathing and attention signals can modulate voxels throughout the brain [24]. IC3 is also an artifact as it consists of regions from cerebrospinal fluids and is a result of cardiac pulsatility artifacts [28]. Removing these two artifacts does not change the performance of the model. The other components consist of some common default mode networks (regions that are shown to be active during rest [53]) as well as some new resting state networks chosen by the algorithm from the ADHD-200 dataset. Some of the resting state networks found are consistent with [15]. For example, IC5 is the resting state network for peristriate area, and lateral and superior occipital gyrus, which are areas related to visual cortex and might represent spontaneous brain activities like day-dreaming [25, 27]. It is easy to connect this component to ADHD as ADHD patients are more likely to experience mind-wandering [48]. IC6 captured connectivity in the frontal and occipital lobe (responsible for planning and many areas of vision respectively). This area is very important for ADHD as well, as ADHD patients may suffer from lack of effective planning [30]. An interesting observation is that IC24 consists of pons (responsible for eye movement, sleep, and many other vegetal and automatic functions) regions and temporal lobe (for sensory processing, memory formation and higher order association processing). Their usefulness in prediction suggests that some of the physical signals captured by fMRI are also indicative of a disease state.

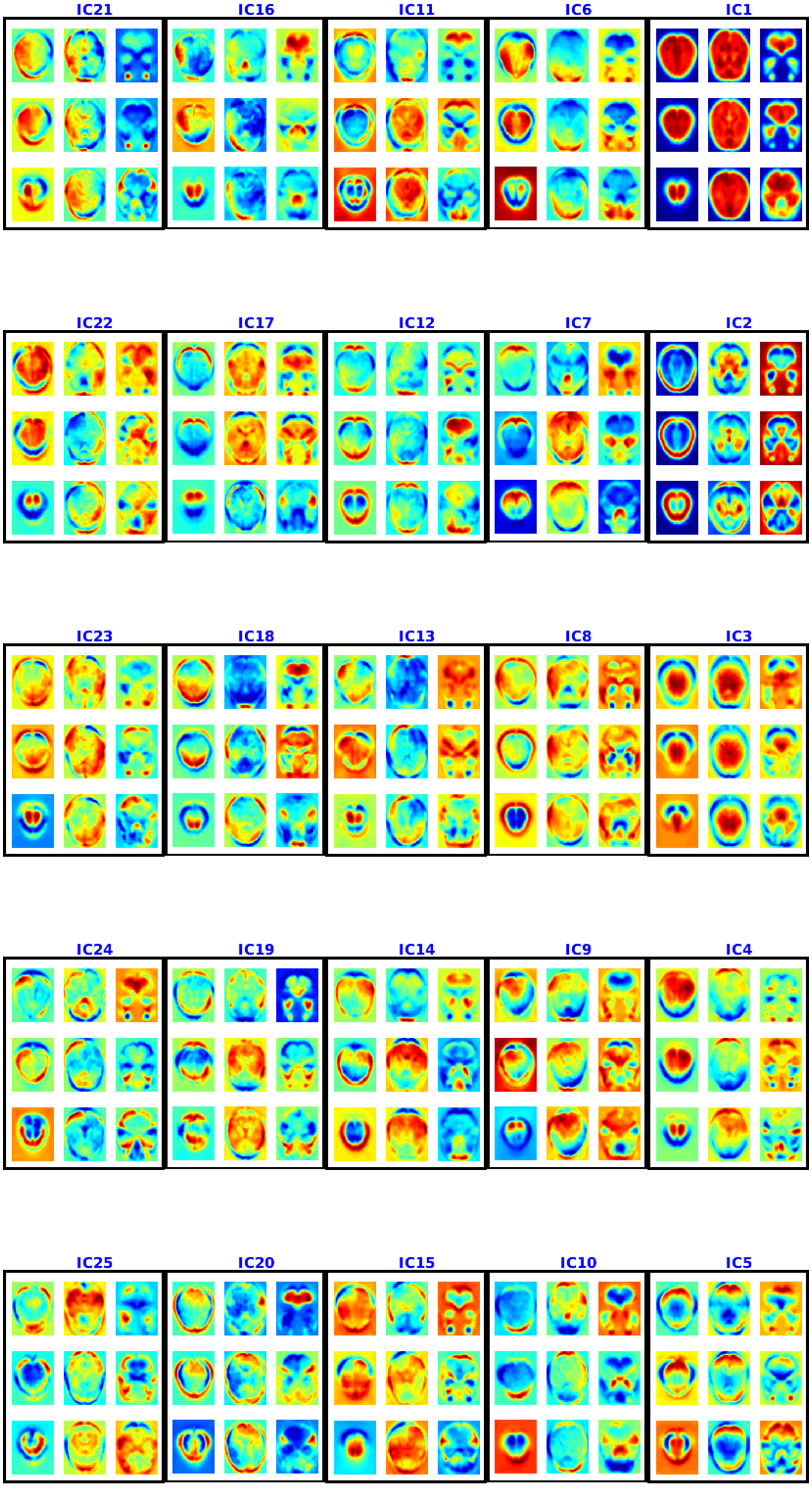


Figure 3.7: Spatial maps or components 1-25 for ADHD-200 using NSD. Each component is shown in a box and 9 axial slices are shown. The colorbar is same as Fig. 3.1

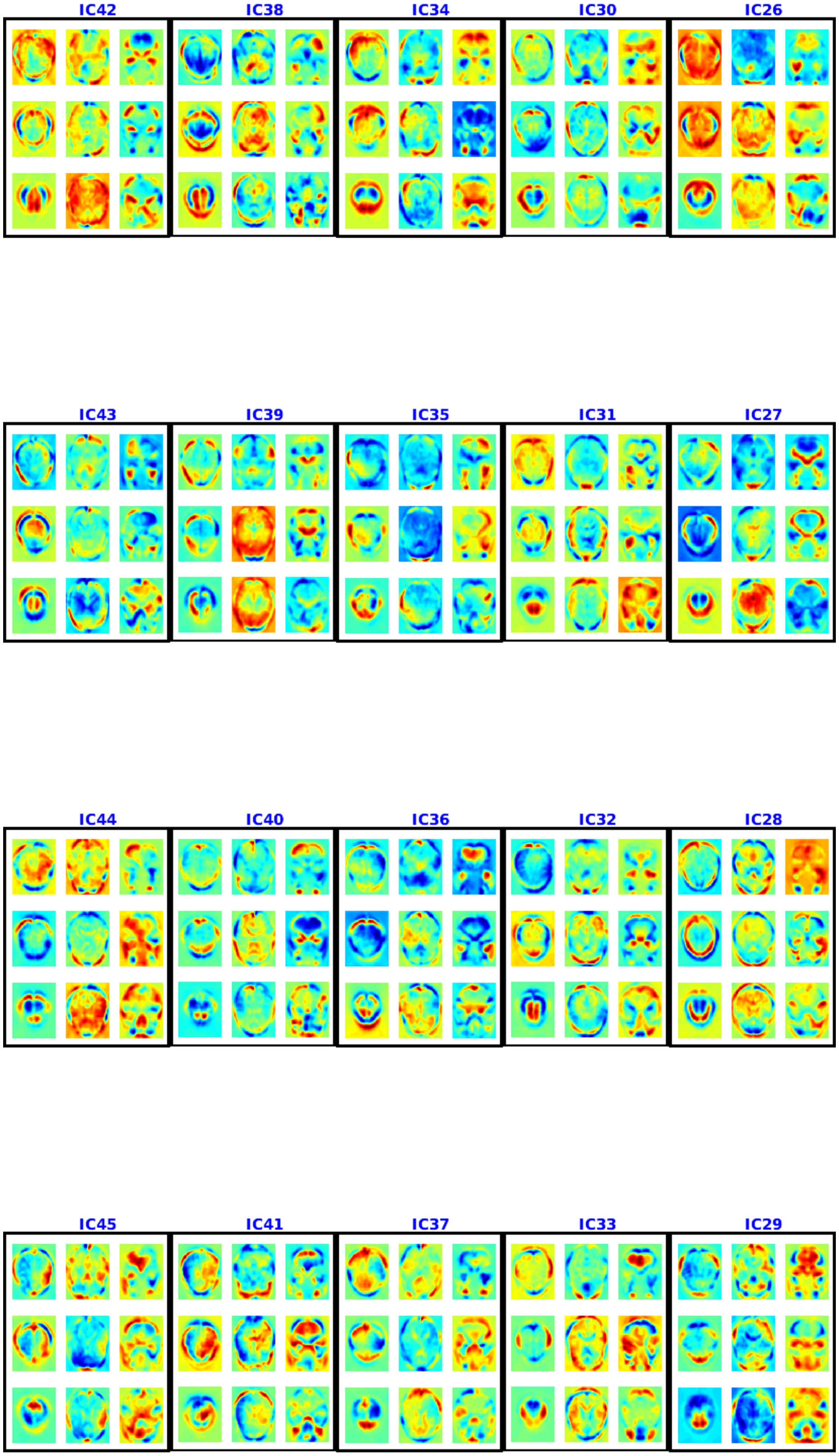


Figure 3.8: Spatial maps or components 26-45 for ADHD-200 using NSD. Each component is shown in a box and 9 axial slices are shown. The colorbar is same as Fig. 3.1

3.3.2 Experiments Using ABIDE Data

Model 5-Fold Cross Validation Training Accuracy

The 5-fold cross-validation accuracy of the model along with the range have been shown in Table 3.5 . The best performing model has a cross-validation accuracy of 0.6225 over baseline 0.5157 (p = 4.9460e-10)

The results for cross validation accuracy for all the source decomposition methods is shown in Fig. 3.9.

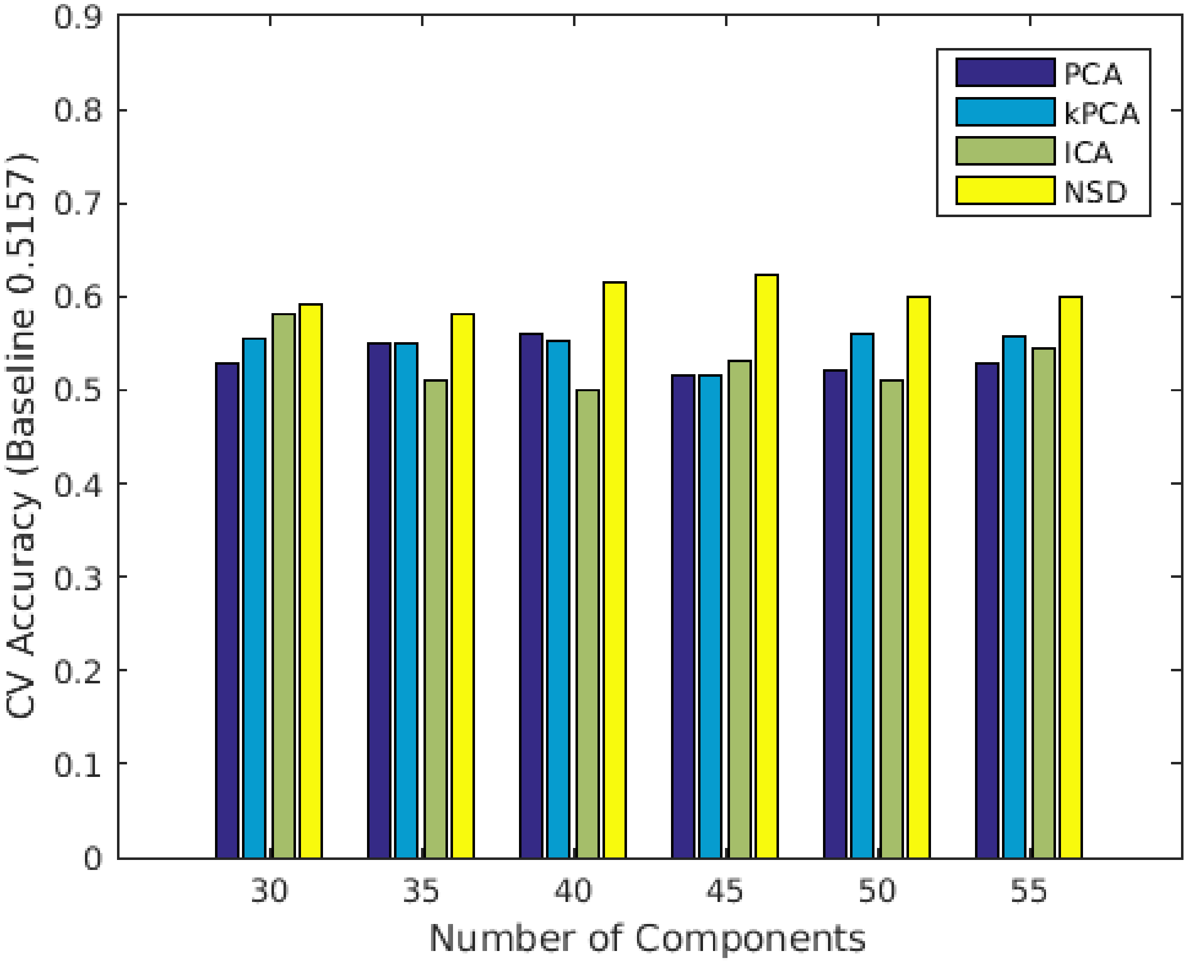


Figure 3.9: 5-fold cross validation results of Autism classification

Model Hold-out Test Accuracy

The accuracy on the hold-out data was 0.6233. This result is statistically significant (p = 0.01) and 2.33% higher with respect to previous best result [10]. The specificity, sensitivity and Jstat are given by 0.6768, 0.5533 and 0.2301 respectively.

Table 3.5: 5-fold cross-validation results for Autism classification using different number of independent components

Number of ICs (SVM Gamma)	5-Fold CV Accuracy	5-Fold CV STD Dev
30(51.2)	0.5925	0.0158
35(51.2)	0.5825	0.0049
40(51.2)	0.6159	0.0078
45(51.2)	**0.6225**	0.0210
50(51.2)	0.6000	0.013
55(51.2)	0.5987	0.0059

Table 3.6: Leave-one-out results for Autism classification using only imaging data

Algorithm	Accuracy	Specificity	Sensitivity	J-Statistics
Our Result	**0.6139**	0.6475	0.5781	0.2256
Nielsen et al [18]	0.60	0.58	0.62	0.20

Leave-one-out Accuracy Comparison

Leave-one-out accuracy of the model was also calculated to compare the results with previous best result. The comparison is shown in Table 3.6.

Spatial Maps from ABIDE data

The spatial maps found using multiple de-correlation are shown in Fig. 3.10 and Fig. 3.11. IC1 is an artifact as it is shared by almost all the voxels [24]. IC3 is also an artifact as it consists of regions from cerebrospinal fluids and is a result cardiac pulsatility artifacts [28]. A deeper investigation into the components shows multiple overlapping components. The components include visual areas (visual cortex, V1 and V2), partial overlapping with some default mode networks (PCC/precuneus, anterior cingulate cortex and frontal lobe) and motor networks. These components are informative: see the cross-validation, test accuracy on the ABIDE data for autism classification.

3.3.3 Discussion on Results

In our analysis we have used the individual time courses for each component as features for the learner. The separation model is $X^i_{[T\times V]} \approx A_{i[T\times I]} \times S_{[I\times V]}$ where rows of S are estimated spatial map and columns of A are corresponding estimated time courses. The time course component $A_{:,i}$ correspond to weighting of the component $S_{i,:}$. A smaller component weight $A_{:,i}$ will correspond to lower contribution of the component to the whole scan and indicate hypo-connectivity. Likewise higher weights for a component will represent

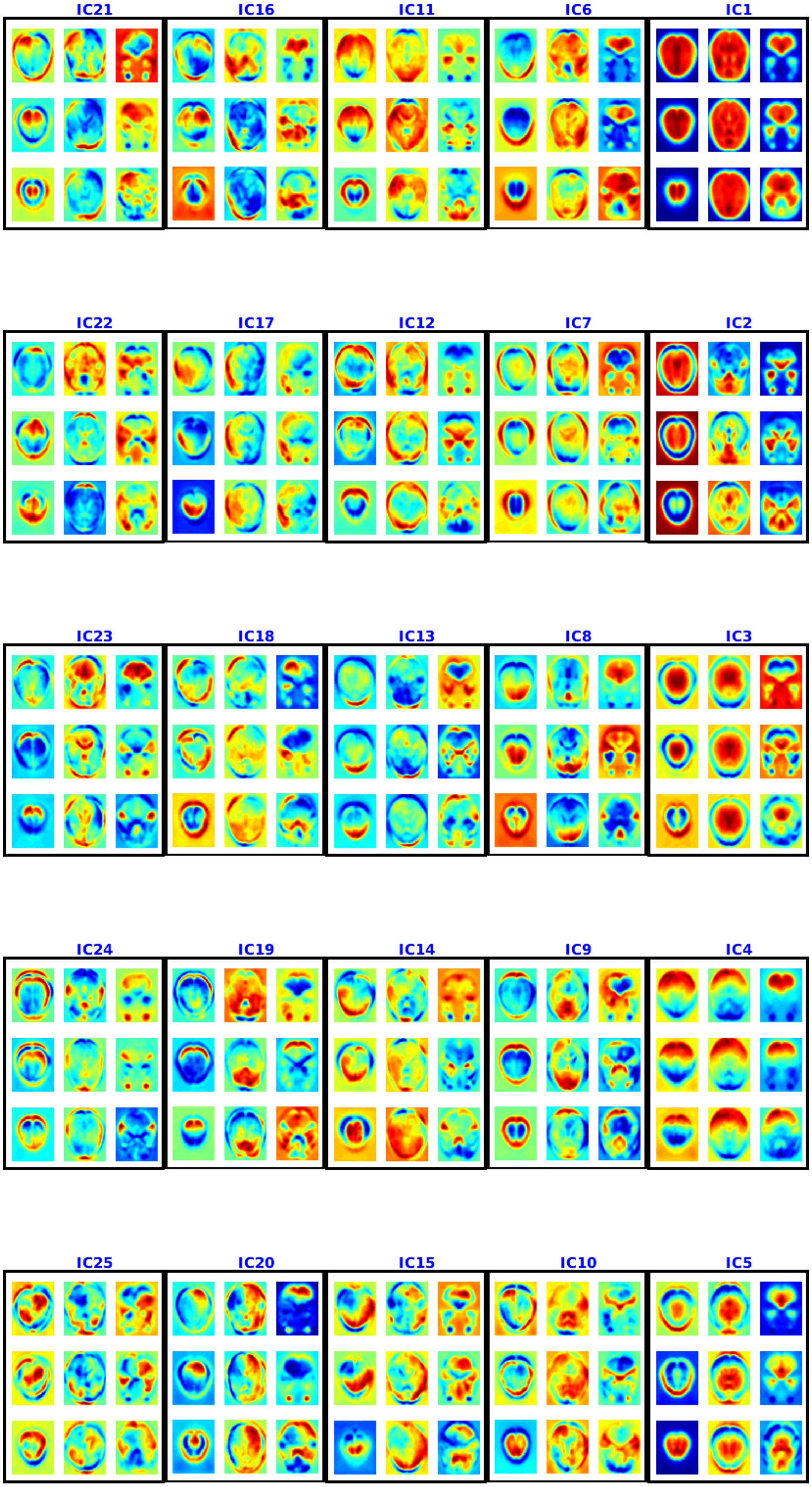


Figure 3.10: Spatial maps or components 1-25 for ABIDE using NSD. Each component is shown in a box and 9 axial slices are shown. The colorbar is same as Fig. 3.1

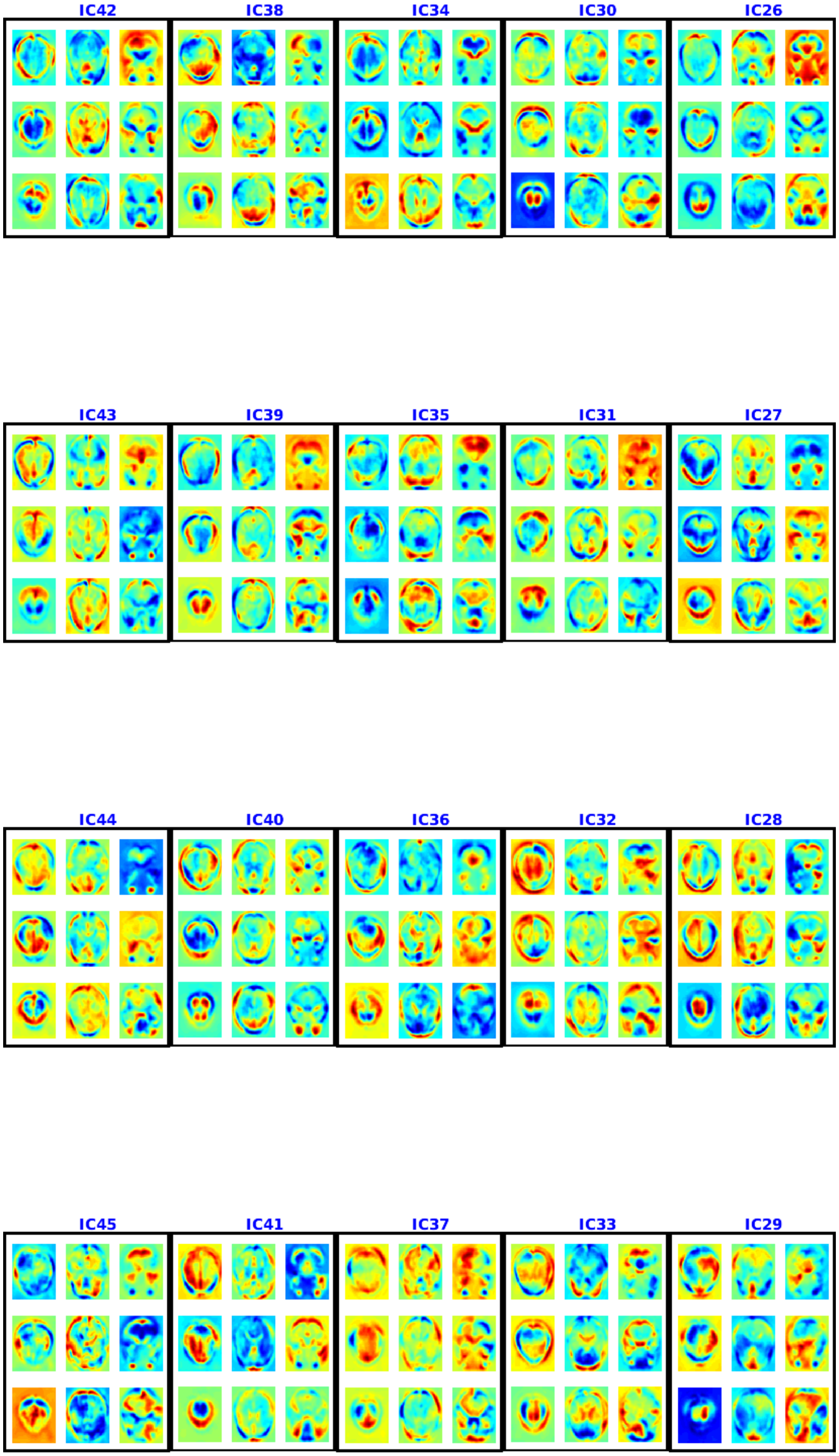


Figure 3.11: Spatial maps or components 26-45 for ABIDE using NSD. Each component is shown in a box and 9 axial slices are shown. The colorbar is same as Fig. 3.1

hyper-connectivity.

For ADHD, different studies have reported different pathological changes in brain [21]. Tian et al. [22] showed a higher level activity in sensory cortex. Whereas using a similar method, Castellanos et al. [23] conceptualized a lower connectivity between anterior cingular cortex, precuneus and posterior cingulate cortex. In our studies, we found differences in group level mean connectivities for ADHD cases for visual and default mode components corresponding to regions denoted in [21, 23] ($S =$ IC4, IC7, IC9, IC28, IC36, IC38, IC42, IC43). We compared corresponding group level mean weights for Healthy and ADHD patients. For IC$i \in S$, corresponding weights $A_{:,i}$ are reduced for ADHD patients in a total of 42, 42, 42, 42, 40, 41, 43, 36 time points out of total number of time points 91. The reduced differences for were statistically significant ($p \leq 0.05$) for 21, 20, 20, 21, 19, 19, 22, 18 number of time points. This means, the patients suffering from ADHD might have a combination of hyper-connective or hypo-connective brain depending on time points.

For Autism, the main connectivity loss is noted in frontal lobe and other cortical areas [19] [20]. In our analysis this corresponds to IC4, IC6, IC7, IC8, IC15, IC16, IC29, IC42, IC41. In all these cases, similar to ADHD, we compared corresponding group level mean weights for Healthy and Autism patients. The weights are reduced for autism patients in a total of 51, 57, 50, 57, 53, 53, 52, 53, 52 time points whereas total number of time points is 91. Also these differences were significant with ($p \leq 0.05$) for 21, 20, 20, 23, 19, 23, 20, 18, 19 number of time points. IC3 predominantly represents CSF ventricle in brain. Our study shows that this component does not have much effect between healthy and autistic brains as there is group level differences ($p \leq 0.05$) in only 18 time points − i.e. only 0.2 of the whole time points of IC3 are significantly different.

3.4 Method 4 (Multi-modal Features for Prediction of Psychiatric Diseases)

5-fold cross validation and test set accuracy results are shown in Tables 3.7 and 3.8.

Table 3.7: 5-fold cross validation results for ADHD classification using features from structural and functional scans

5-Fold CV Accuracy	Hold-out Accuracy
0.6892	0.6725

Table 3.8: 5-fold cross validation results for Autism classification using features from structural and functional scans

5-Fold CV Accuracy	Hold-out Accuracy
0.6312	0.6431

For both of these datasets (ADHD-200 and ABIDE), these accuracy values are the best known using only imaging data. For the ADHD-200 hold-out test, the specificity, sensitivity and Jstat are 0.8510, 0.4545, 0.3055 respectively. For the ABIDE hold-out test, specificity, sensitivity and Jstat are 0.6832, 0.6, 0.2832 respectively.

3.4.1 Discussion

As shown in [10], texture based features from fMRI and MRI scans can be predictive of psychiatric diseases. Our model derives the texture based features from MRI and combines them with resting state information from fMRI to produce a strong predictor.

Comparison with Previous results

Though there is a significant improvement of the prediction results compared to previous works, the model is able to increase the prediction accuracy only by 4.65% for ADHD and 4.31% for Autism compared to the previous works on ADHD/Autism prediction. There are several reasons for the lackluster performance. We can hypothesize that the resting state network structures are not significantly different between ADHD/Autism patients and healthy controls. As can be seen from the previous section, almost half of the total number of time points for each spatial component had no statistical differences between healthy and ADHD/Autism positives. Also it is later shown (Fig. 3.12) that the pre-processing step does not remove all the site dependent artifacts for the fMRI data.

Effect of Unbalanced Data for ADHD-200

Here we discuss effect of unbalanced data for ADHD classification. In the case of autism classification from ABIDE data, the class labels (healthy vs. autism) are almost equally distributed (baseline 0.5157), hence we did not use balanced data for training and testing. Instead of using the balanced training set for ADHD classification, had we used the unbalanced training set (with baseline 0.6372), the training accuracy for the model rises to 0.715 (std 0.043 and p=1.9388e-06). However the hold-out test accuracy drops to 0.6257. This result shows that balancing the training set may be a good way to improve the classifier prediction accuracy. One drawback of this method is it may not be representative of the population as only 11% of children 4-17 years of age have been diagnosed with ADHD in United States as of 2011 [6].

Reliability of Multisite Dataset

Although pre-processing was applied to the data before applying any machine learning step, the datasets suffer significantly from site dependent variations. Both ADHD-200 and ABIDE data were passed through careful experimental control and quality assurance checks. Even

[6] `http://www.cdc.gov/nchs/fastats/adhd.htm`

after that the site dependent variations have a heavy impact on the data. For ADHD-200, Fig. 3.12 shows the first two principal components where data from each site is shown using one color (independent of ADHD vs control label). Ideally we should see data from different sites intermixed together but instead we see clusters corresponding to different sites.

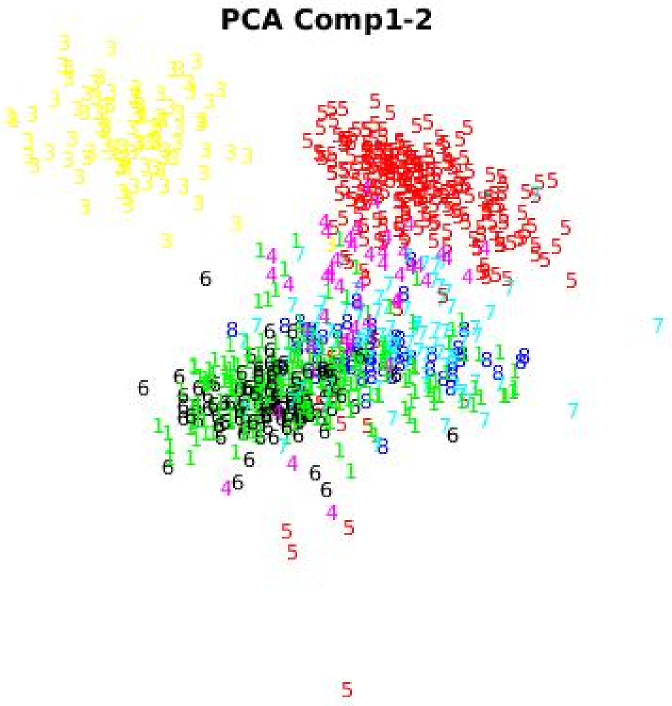


Figure 3.12: PCA Component 1-2 for different sites. Each number represents one site. Here, x-y axes are PCA component 1 and PCA component 2 respectively.

Most of the papers working on ADHD-200/ABIDE data suffer from this fatal caveat. The first few principal components among the fMRI scans account for most site dependent variations. However if we attempt to learn a model based on removing first few components to make the data more homogeneous, the model loses its predictability, indicating that the first few important principal components are important features for disease classification too. We can use the fMRI ICA features to learn a classifier that is able to predict site from which the subjects come from, with 92% accuracy. Hence we hypothesize that ADHD/Autism predictability is also interlinked with site dependent fMRI scan features − i.e. there are some common features that can predict both the site and the disease. These two factors should be decoupled before we can do a true analysis of ADHD/Autism predictability using fMRI features. However, our analysis is an important step towards an automated generalized prediction model for psychiatric disease detection.

Chapter 4

Conclusion

The development of automatic ADHD/Autism diagnostic algorithms from MRI/fMRI data is a challenging task. The application of statistical pattern recognition algorithms to this problem currently yields only moderately good results, rendering these classification systems unfit for deploying in the health-care industry. However, much research is being done to improve these results and search for discriminative features for classifying ADHD/Autism amongst the plethora of voxel values present in structural (MRI) and functional neuroimages (fMRI).

In this dissertation, we derived a novel algorithm for combining structural and functional features using 3D texture based and independent component analysis of the whole 4-D fMRI scan, which can then be used for classification. We also explored different representation of brain functional connectivity useful for differentiating Healthy vs Psychiatric patients. Our results indicate that combining multimodal features (both MRI and fMRI) yields moderately good classification accuracy for ADHD/Autism, which is an important step towards computer aided diagnosis of these psychiatric diseases.

Still, there is much work to be done in this area. For example, source separation based on deep belief networks [29] can be investigated in this vein. Using other multimodal (Diffusion Tomographic Imaging, Electroencephalogram) features also be used for prediction. Moreover we identify site independent characteristics for any feature extraction an important challenge of the problem and future efforts should be directed in this vein.

Bibliography

[1] Calhoun VD, Adali T, Pearlson GD, Pekar JJ (2001b). A method for making group inferences from functional MRI data using inde- pendent component analysis. Human Brain Mapping 14:140151

[2] Calhoun VD, Adali T, Pearlson GD, Pekar JJ (2002). Erratum: A method for making group inferences from functional MRI data using independent component analysis. Human Brain Mapping 16:131.

[3] Eloyan, Ani, John Muschelli, Mary Beth Nebel, Han Liu, Fang Han, Tuo Zhao, Anita D. Barber et al. Automated diagnoses of attention deficit hyperactive disorder using magnetic resonance imaging. Frontiers in systems neuroscience 6 (2012).

[4] Li S, Eloyan A, Joel S, Mostofsky S, Pekar J, Bassett S S, Caffo B. (2012). Analysis of Group ICA-Based Connectivity Measures from fMRI: Application to Alzheimer's Disease. Plos One. 7:11. e49340.

[5] Smith, Stephen M., et al. Temporally-independent functional modes of spontaneous brain activity. Proceedings of the National Academy of Sciences 109.8 (2012): 3131-3136.

[6] Liu, Xiao, and Jeff H. Duyn. Time-varying functional network information extracted from brief instances of spontaneous brain activity. Proceedings of the National Academy of Sciences 110.11 (2013): 4392-4397.

[7] Parra L., Spence C. 2000. Convolutive Bilnd Separation of Non-Stationary Sources *IEEE transaction on Speech and Audio Processing* Vol.8 NO. 3: 320–327.

[8] Souloumiac, Antoine. Blind source detection and separation using second order non-stationarity. Acoustics, Speech, and Signal Processing, 1995. ICASSP-95., 1995 International Conference on. Vol. 3. IEEE, 1995.

[9] Cecchi, G., I. Rish, B. Thyreau, B. Thirion, M. Plaze, and M. Paillere-Martinot. 2009. Discriminative network models of schizophrenia. *Advances in Neural Information Processing Systems* 22 (2009): 252–60.

[10] Ghiassian S., Greiner R., Jin P., Brown M. 2014. Learning to Classify Psychiatric Disorders based on fMR Images: Autism vs Healthy and ADHD vs Healthy. *In Proceedings of 3rd NIPS 2013 Workshop on Machine Learning and Interpretation in NeuroImaging* (2014).

[11] Nielsen, J. A., Zielinski, B. A., Fletcher, P. T., Alexander, A. L., Lange, N., Bigler, E. D., Anderson, J. S. (2013). Multisite functional connectivity MRI classification of autism: ABIDE results. Frontiers in human neuroscience, 7.

[12] Li Y., Adal T.,Calhoun V., 2007. Estimating the Number of Independent Components for Functional Magnetic Resonance Imaging Data. In *Human Brain Mapping*, 28:12511266.

[13] Dai, D., Wang, J., Hua, J., He, H. (2012). Classification of ADHD children through multimodal magnetic resonance imaging. Frontiers in systems neuroscience, 6.

[14] Sidhu, G. S., Asgarian, N., Greiner, R., Brown, M. R. (2012). Kernel Principal Component Analysis for dimensionality reduction in fMRI-based diagnosis of ADHD. Frontiers in systems neuroscience, 6.

[15] Damoiseaux, J. S., Rombouts, S. A. R. B., Barkhof, F., Scheltens, P., Stam, C. J., Smith, S. M., Beckmann, C. F. (2006). Consistent resting-state networks across healthy subjects. Proceedings of the national academy of sciences, 103(37), 13848-13853.

[16] Brown, G. G., Mathalon, D. H., Stern, H., Ford, J., Mueller, B., Greve, D. N., Network, F. B. I. R. (2011). Multisite reliability of cognitive BOLD data. Neuroimage, 54(3), 2163-2175.

[17] Jovicich, J., Czanner, S., Greve, D., Haley, E., van der Kouwe, A., Gollub, R., Dale, A. (2006). Reliability in multi-site structural MRI studies: effects of gradient non-linearity correction on phantom and human data. Neuroimage, 30(2), 436-443.

[18] Nielsen, J. A., Zielinski, B. A., Fletcher, P. T., Alexander, A. L., Lange, N., Bigler, E. D., Anderson, J. S. (2013). Multisite functional connectivity MRI classification of autism: ABIDE results. Frontiers in human neuroscience, 7.

[19] Cherkassky, V. L., Kana, R. K., Keller, T. A., Just, M. A. (2006). Functional connectivity in a baseline resting-state network in autism. Neuroreport, 17(16), 1687-1690.

[20] Just, M. A., Cherkassky, V. L., Keller, T. A., Kana, R. K., Minshew, N. J. (2007). Functional and anatomical cortical underconnectivity in autism: evidence from an FMRI study of an executive function task and corpus callosum morphometry. Cerebral cortex, 17(4), 951-961.

[21] Konrad, K., Eickhoff, S. B. (2010). Is the ADHD brain wired differently? A review on structural and functional connectivity in attention deficit hyperactivity disorder. Human brain mapping, 31(6), 904-916.

[22] Tian L, Jiang T, Wang Y, Zang Y, He Y, Liang M, Sui M, Cao Q, Hu S, Peng M, Zhuo Y (2006). Altered resting-state functional connectivity patterns of anterior cingulate cortex in adolescents with attention deficit hyperactivity disorder. Neurosci Letter,400:3943.

[23] Castellanos FX, Margulies DS, Kelly C, Uddin LQ, Ghaffari M, Kirsch A, Shaw D, Shehzad Z, Di Martino A, Biswal B, Sonuga-Barke EJ, Rotrosen J, Adler LA, Milham MP (2008). Cingulate-precuneus interactions: A new locus of dysfunction in adult attention-deficit/hyperactivity disorder. Biol Psychiatry 63:332337.

[24] Power, J. D., Mitra, A., Laumann, T. O., Snyder, A. Z., Schlaggar, B. L., Petersen, S. E. (2014). Methods to detect, characterize, and remove motion artifact in resting state fMRI. Neuroimage, 84, 320-341.

[25] Wang, K., Jiang, T., Yu, C., Tian, L., Li, J., Liu, Y., Li, K. (2008). Spontaneous activity associated with primary visual cortex: a resting-state FMRI study. Cerebral cortex, 18(3), 697-704.

[26] Deshpande, G., Wang, P., Rangaprakash, D., Wilamowski, B. (2015). Fully Connected Cascade Artificial Neural Network Architecture for Attention Deficit Hyperactivity Disorder Classification From Functional Magnetic Resonance Imaging Data.

[27] Christoff, K., Gordon, A. M., Smallwood, J., Smith, R., Schooler, J. W. (2009). Experience sampling during fMRI reveals default network and executive system contributions to mind wandering. Proceedings of the National Academy of Sciences, 106(21), 8719-8724.

[28] Dagli, M. S., Ingeholm, J. E., Haxby, J. V. (1999). Localization of cardiac-induced signal change in fMRI. Neuroimage, 9(4), 407-415.

[29] Hjelm, R. D., Calhoun, V. D., Salakhutdinov, R., Allen, E. A., Adali, T., Plis, S. M. (2014). Restricted Boltzmann machines for neuroimaging: an application in identifying intrinsic networks. NeuroImage, 96, 245-260.

[30] Barkley, Russell A. Behavioral inhibition, sustained attention, and executive functions: constructing a unifying theory of ADHD. Psychological bulletin 121.1 (1997): 65.

[31] Coates, A., Ng, A. Y., Lee, H. (2011). An analysis of single-layer networks in unsupervised feature learning. In International conference on artificial intelligence and statistics (pp. 215-223).

[32] Gupta, A., Ayhan, M., Maida, A. (2013). Natural image bases to represent neuroimaging data. In Proceedings of the 30th International Conference on Machine Learning (ICML-13) (pp. 987-994).

[33] Bourlard, H. and Kamp, Y. Auto-association by multilayer perceptrons and singular value decomposition. Biological Cybernetics, 59(4):291294, 1988

[34] Olshausen, B. A. Sparse codes and spikes. In Probabilistic Models Of The Brain: Perceptron Aand Neural Function, pp. 257272. MIT Press, 2001

[35] Lecun, Y. and Bengio, Y. Convolutional Networks for Images, Speech and Time Series, pp. 255258. The MIT Press, 1995

[36] Peng, H., Long, F., Ding, C. (2005). Feature selection based on mutual information criteria of max-dependency, max-relevance, and min-redundancy. Pattern Analysis and Machine Intelligence, IEEE Transactions on, 27(8), 1226-1238.

[37] Donahue, J., Jia, Y., Vinyals, O., Hoffman, J., Zhang, N., Tzeng, E.,Darrell, T. (2013). Decaf: A deep convolutional activation feature for generic visual recognition.

[38] Sermanet, P., Eigen, D., Zhang, X., Mathieu, M., Fergus, R., LeCun, Y. (2013). Overfeat: Integrated recognition, localization and detection using convolutional networks. arXiv preprint arXiv:1312.6229.arXiv preprint arXiv:1310.1531.

[39] Razavian, A. S., Azizpour, H., Sullivan, J., Carlsson, S. (2014, June). CNN features off-the-shelf: an astounding baseline for recognition. In Computer Vision and Pattern Recognition Workshops (CVPRW), 2014 IEEE Conference on (pp. 512-519). IEEE.

[40] Deshpande, G., Wang, P., Rangaprakash, D., Wilamowski, B. (2015). Fully Connected Cascade Artificial Neural Network Architecture for Attention Deficit Hyperactivity Disorder Classification From Functional Magnetic Resonance Imaging Data.

[41] . D. Zeiler and R. Fergus. Visualizing and understanding convolutional networks. CoRR, abs/1311.2901, 2013

[42] M. Oquab, L. Bottou, I. Laptev, and J. Sivic. Learning and transferring mid-level image representations using convolutional neural networks. Technical Report HAL-00911179, INRIA, 2013.

[43] Calhoun VD, Kiehl KA, Pearlson GD. Modulation of temporally coherent brain networks estimated using ICA at rest and during cognitive tasks. Human Brain Mapping. 2008a; 29(7):828. [PubMed: 18438867]

[44] Schmithorst VJ, Holland SK. Comparison of three methods for generating group statistical inferences from independent component analysis of functional magnetic resonance imaging data. J.Magn Reson.Imaging. 2004; 19(3):365368. [PubMed: 14994306]

[45] Raina, R., Battle, A., Lee, H., Packer, B., and Ng, A. Y. Self-taught learning: Transfer learning from unlabeled data. In ICML, pp. 759766, 2007.

[46] Erhardt, E. B., Rachakonda, S., Bedrick, E. J., Allen, E. A., Adali, T., Calhoun, V. D. (2011). Comparison of multisubject ICA methods for analysis of fMRI data. Human brain mapping, 32(12), 2075-2095.

[47] Anderson, A., Douglas, P. K., Kerr, W. T., Haynes, V. S., Yuille, A. L., Xie, J., Cohen, M. S. (2014). Non-negative matrix factorization of multimodal MRI, fMRI and phenotypic data reveals differential changes in default mode subnetworks in ADHD. NeuroImage, 102, 207-219.

[48] Shallice T, Marzocchi GM, Coser S, Del Savio M, Meuter RF, Rumiati RI. Executive function profile of children with attention deficit hyperactivity disorder. Developmental Neuropsychology. 2002;21:4371.

[49] Calhoun, V., Adali, T., Liu, J. (2006, August). A feature-based approach to combine functional MRI, structural MRI and EEG brain imaging data. In Engineering in Medicine and Biology Society, 2006. EMBS'06. 28th Annual International Conference of the IEEE (pp. 3672-3675). IEEE.

[50] Brown MRG, Sidhu GS, Greiner R, Asgarian N, Bastani M, Silverstone PH, Greenshaw AJ and Dursun SM (2012) ADHD-200 Global Competition: diagnosing ADHD using personal characteristic data can outperform resting state fMRI measurements. Front. Syst. Neurosci. 6:69. doi: 10.3389/fnsys.2012.00069

[51] A. Hyvrinen. Fast and Robust Fixed-Point Algorithms for Independent Component Analysis. IEEE Transactions on Neural Networks 10(3):626-634, 1999.

[52] Lai, S. H., Fang, M. (1999). A novel local PCA-based method for detecting activation signals in fMRI. Magnetic resonance imaging, 17(6), 827-836.

[53] Biswal B, Yetkin FZ, Haughton VM, Hyde JS. (1995). Functional connectivity in the motor cortex of resting human brain using echo-planar MRI. Magnetic resonance in medicine, Oct;34(4):537-41.

[54] Souloumiac (1995). Blind source detection and separation using second order nonstationarity. In International Conference on Acoustics, Speech and Signal Processing, volume IEEE 0-7803-2431-5/95, pages 1912-915.

[55] Kawamoto, M., Matsuoka, K., and Ohnishi, N. (1998). A method of blind separation for convolved non-stationary signals. Neurocomputing, 22:157-171.

[56] http://stats.stackexchange.com/questions/114385/what-is-the-difference-between-convolutional-neural-networks-restricted-boltzma

[57] http://202.118.75.4/ma/bss.html